STATISTICAL METHODS
IN
EDUCATION

by
R. P. SINGH VERMA

1969
20.9.64

311
VER

Statistical Methods
IN
EDUCATION
FOR
B. T. & B. Ed. CLASSES

By

R. P. SINGH VERMA *M. Sc.*

Assistant Professor
Department of Statistics
JAT VEDIC (POST-GRADUATE) COLLEGE,
BARAUT (MEERUT).

Published by
RAMA PUBLISHING HOUSE
BARAUT (MEERUT)

Published by
RAMA PUBLISHING HOUSE
Baraut [Meerut].

311
VER

Price Rs. 6·50 only

ALL RIGHTS RESERVED WITH THE AUTHORS

Printed at
SINGHAL PRESS,
Meerut.

To

M. Pratap.

S.C.E.R.T., West Bengal
Date..............................
Acc. No. 1969

Preface

The present text "*Statistical Methods in Education*" has been written to meet the requirements of the students preparing for B. Ed. and B. T. examinations in which statistics is set as a subsidiary subject and is a compulsory part of the paper of '*Educational Psychology*'. The book will be found useful to those who are beginners and want to have an *Elementary Idea of Statistics* and *Basic Statistical concepts*.

Although in an extensive subject of study like statistics, it is quite inevitable that no single text, however bulky it may be, can adequately cover up the whole range of the topics required in the text. Yet an attempt has been made in this book to cover the courses prescribed for the examinations of B.Ed. and B.T. Many of these students have limited knowledge of *Mathematics* and so the mathematical investigations have, in all cases, been rendered as simple and lucid as possible, and have been developed so as to direct attention to the statistical aspect of the subject. A number of illustrative examples which may readily be performed are fully described in each chapter, and numerous questions, mostly selected from universities and public examination papers, are appended. The present book, I believe, has a large number of tables, graphs, diagrams and exercises of all types firstly to satisfy the need of those for whom it is intended and secondly to illustrate the theory amply.

The reader is expected to be familiar with Arithmatic of school level and a limited knowledge of algebraic symbols. The writer has purposely avoided proving the formulae, lest it should unnecessarily burden the minds of the students.

It is particularly important that the reader should not form the impression that *Statistical Methods* contain a series of incomprehensible formulae to be applied indiscriminately to any available data. However care should be taken to grasp and appreciate the ideas and principles underlying the method learnt.

The whole book is divided into four big chapters namely—*Classification and Tabulation; Graphs and Diagrams; Measures of*

Central Tendency [averages] and Dispersion, Correlation and Regression. The diagrams in chapter second and regression in chapter fourth are introduced in the text to give a better understanding to the reader regarding the *Basic Statistical Concepts* and an *Elementary Idea of Statistics.*

Author of a text book is always indebted to all the earlier books on the subject and therefore my indebtedness to original memoirs and all the earlier books on the subject is acknowledged. I express my manifold deep gratitude to all the publishers and writers whose book I have often consulted.

The author is indebted to all the colleagues for their critical comments and valuable suggestions which have undoubtedly improved the book. In particular, I consider it my noble duty to express my deep gratitude to reverend Gurujee Prof. Mahendra Pratap, Head of the Statistics Deptt,, J. V. College, Baraut, and Prof. S. P. Singh, Assistant Professor of Statistics, J. V. College, Baraut (Meerut) who have given their valuable assistance at all stages and who have placed their precious collections of a few problems given in the text. I must, however, here return my thanks to Prof. A. C. Sharma and B. S. Gupta Assistant Professors in B. Ed., D. J. (Post-Graduate) College, Baraut, for their kind inspiration to write the text. Finally, my sincere thanks are due to Prof. Fauran Singh and H P. Singh, Assistant Professors of Statistics, J. V. College, Baraut, for their courteous assistance and advice while the sheets have been passing through the press.

Inspite of my best efforts, a number of misprints and mistakes are likely to have crept in a book of this type. I shall be grateful to any notice of these errors and misprints and for any suggestion to improve the text.

1st August, 1965.
Baraut.

Verma

Introduction

The word '*Statistics*' is used in two sensenses. Sometimes it is used, as a sigular and sometimes s plural of '*Statistic*'. When it is used in the former sense, refers to the subject of Statistics. or the theory of Statistics. The later use of the word '*Statistics*' refers to numerical statements of facts. For example, the height of individuals may be described by expressions such as 62″, 66″ etc.; the numerical statements of facts. If we describe the individuals as tall, short, medium etc., the statement is not numerical.

Statistics as referred in the former sense in the previous paragraph, has been defined differently by different writers. The fundamental reason for the differences in these definitions is that Statistics has been an ever growing subject. In the beginning, *Statistics* was considered as a state-craft only. In modern times it has pervaded almost in all branches of human knowledge. Therefore the former definitions of '*Statistics*' are not so comprehensive as those given by modern writers.

Bowley first defined *Statistics* as "*the science of counting and the science of averages*". Afterwards he gave another definition of *Statistics* as "*the science of measurement of social organism regarded as a whole in all its manifestations.*" King defined it as "*the method of judging collected natural or social phenomena from the results obtained by the analysis of an enumeration or collection of estimates.*" These are some of the definitions given by earlier writers on the subject, but none of these definitions correctly refers to the present scope of Statistics.

Statistics is loosely referred to two separate concepts (i) descriptive statistics, (ii) statistical methods. Descriptive statistics means the collection and summarization of numerical data. The second concept is far more important as it deals with the various methods of representing, comparing, analysing and interpreting numerical data. Thus *Statistics* may be defined as *the scientific exposition of statistical methods.*

Statistical methods in modern sense range from the mere recording and tabulation of numerical data to *"subtle process of inductive reasoning based on the mathematical theory of probability."* The present text deals with the simpler methods which involve little more than elementary arithmatic. The most attractive developments in technique which have made the science of statistics so important to the modern research worker are relatively recent *i.e.* mainly in the last forty years. In industry, statistical methods have been used to control the quality of the product. This technique is known as *Quality Control*. Another important use of statistical methods is in Agriculture and is known as the *Design of Experiments*. Statistical methods have also been abundantly used in Biology, Psychology, Education and Engineering in modern times.

Statistical methods have proved their usefulness in business and science. The need for obtaining statistical data in the smooth functioning of any business concerned increases along with size. The bigger the undertaking the greater is the need for Statistics. No business, large or small public or private can thrive without the aid of Statistics. A manufacturer must know in advance the quantity to be produced, the number of workers, the amount of raw material etc. *i.e.* the details of a production plan. Statistical data and Statistical methods render valuable services in the proper understanding of economic problems. Economic problems are always capable of being reduced to numerical facts *e.g.* volume of trade (import and export), output of industries (manufacturing and mining etc.), agriculture (wages, livestocks, area and yield of crops, agricultural prices and forest) etc. These numerical figures are affected by innumerable causes, and items vary according to time and place Thus the study of economic problems is particularly suited to statistical approach. This approach leads to its correct description. Five year plans have made considerable use of this treatment. Apart from economic policy, the development of economic theory based on inductive reasoning is also facilitated by the use of Statistics. Its importance in the field of agriculture, science, and even in technology is in the increase. The importance of Statistics is also reflected in the large number of sciences which set a separate paper for study. For this purpose, the student is required to learn the basic ideas and principles of the subject and it is with such fundamentals that this book is concerned.

It should be understood that it cannot be applied to all kinds of phenomena. The essential condition for the application of statistical methods is that the facts should be capable of being measured quantitatively. Health, poverty, intelligence are the examples where direct numerical measurements are not possible. But the steps are taken to have an indirect measurement such as death rate, longevity of life etc. to make a judgement on the health of the people Intelligence of students may be compared on the basis of marks obtained by them in a class test. Another defect is that statistics attaches no importance to individual items. Therefore it is capable of studying those problems only, where group characteristic come in. One more defect of statistics lies in the fact that statistical laws are not correct like the mathematical laws. By observing a limited number of items (*i.e.* by taking a sample) we infer about the characteristic of the whole population. Sometimes statistics is used to draw wrong conclusions. There is an oft quoted saying that "*there are three kinds of lies—namely lies, damned lies and statistics.*'! Another saying is that "*with Statistics anything can be proved.*" If we make an analysis of such kinds of misgivings, we are apt to conclude that a number of persons exists in this world who is selfish and figures are capable of being manipulated. Therefore the useful application of statistics depends upon the ability, skill and special experience and honesty of the statistician combined with the unbiasedness of the enumerator or investigator.

Contents

Introduction : Meanings and definitions of statistics: statistical methods and their usefulness; scope of statistics; defects of statistics; acknowledgements.

Chapter I. *(Classification and Tabulation)* :

Definition; advantages and basis of classification; types of classification; class interval and frequency table: steps for frequency distribution; methods o classification; types of tabulation. ...1-

Chapter II. *(Graphs and Diagrams)* :

Use and limitations of graphs and diagrams; graphical representations; histograms, frequency polygons and curves; cumulative frequency curves (ogive) and histograms; types of diagrams *i.e.* simple, multiple, sub-divided, Percentage sub divided and bilateral bar diagrams; pie/circle and rectangular diagrams. ...27—88

Chapter III. *(Measures of Central Tendency and Dispersion)* :

Measures of central tendency (averages), definitions: advantages of mean, mode and median; methods of computing the averages ; their properties; quartiles and their computations; Measures of dispersion *i.e.* range, quartile deviation, mean deviation, standard deviation and their computations; coefficient of variation, standard error and variance, types o curves. ...8

Chapter IV. *(Correlation and Regression)* :

Meaning of correlation; definition and direction of correlation; degree of correlation *i.e.* perfect limited degree of +ve, no correlation, limited and of −ve and perfect −ve correlation; measur correlation *i.e.* scatter diagram, Person's coefficient of

correlation, regression line; range of correlation coefficient; definition of regression coefficient, regression lines of Y on X and of X on Y; relation between correlation coefficient and regression coefficients, tests of significance of the regression coefficients. ...149—184

Appendix (1). (*Tables*) :

Values of 't', χ^2 and F distributions for 5% and 1% probability points.

Appendix (2). (*A. U. Examination Papers*) :

Unsolved papers of Agra University Examinations of B. T. & B. Ed.

Chapter 1

Classification & Tabulation

The raw statistical data which is collected in the course of an inquiry, usually consists of a series of readings and measurements. The original form of the collected data is very complex & voluminous and so it is very difficult to draw any inference from it as such. Thus, it is necessary to condence the data and present it in a systematic way. It is done by the 'Classification & Tabulation" of the data. By these processes, the volume and complexity of the data are reduced, which make the analysis and interpretations of the data easier.

Classification:—It is the process of arranging the individuals or items into groups or classes according to their similarities with respect to some variable character.

The Advantages of classification:—

(1) The whole data is divided into a number of classes as the items having resemblances with respect to certain character are put together in one class. Thus, it indicates the point of similarity and dis-similarity.

(2) The classification reduces the volume of the data which helps in forming the mental picture of the data.

(3) It brings into light those important informations which are liable to be ignored without classifying the data.

(4) It prepares the ground for comparisons and inferences. Because the classification of the data is not only sufficient or the comparisons and interpretations but it also helps in the tabulation of the data. Thereafter, the comparisons and the

interpretations on the data are possible.

(5) The classification provides the orderly arrangement of the items of the data.

Basis of classification:

The classification of the data is based upon the characteristic possessed by the items contained in the data. There are two types of characteristics:—

(i) Descriptive (ii) Numerical.

(i) **Descriptive-characteristic**: Under this type of characteristic, the quantitative measurements are not possible but only the presence or absence of a particular characteristic is observed in the terms available in the data. For example, sex, literacy, unemployment and blindness etc are the D*escriptive or qualitative characteristic* of the items.

Classification according to Attribute: The classification of the given data according to a descriptive characteristic is known as the classification according to attribute. Here only the presence or absence of a quality (attribute) forms the basis of classification

The classification can further be divided into two kinds;—

(a) Simple classification or classification by dichotomy,

(b) Manifold classification.

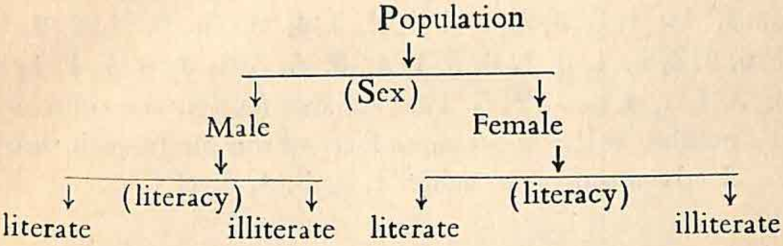
(a) **Simple Classification**: The kind of classification is termed as Simple classification where only one attribute is taken as the criterion of classification. Here, the whole data is divided into two classes:—

(1) Having those items which posses the attribute quality),

(2) Having those items which do not possess the attribute.

For example; the population of a country can be divided into two classes according to the attribute 'Unemployment' i.e. unemployed and employed or according to the attribute 'sex' i,e. male and female etc.

(b) **Manifold Classification**. The kind of classification is termed as Manifold classification where more than one attribute is taken as the criteria of classification. Here the whole data is divided into more than two classes. For example, the population of a country can be divided into four classes according to two attributes 'sex' & 'literacy' as shown in the diagram given below.



Thus the sub-divisions of the population are—

(1) literate male (2) Illiterate male (3) literate female (4) Illiterate female

Note:—

In classification according to attribute, it is necessary that the classes should be clearly defined before starting the actual work of classification i. e. a boundary should be set between the classes. As for example, in case of literacy we must decide in advance to whom we shall call the literate and to whom the illiterate before starting the work of classification of the data.

(ii) **Numerical Characteristic**: Under this type of characteristic, the quantitative measurements are possible. As prices, income, age, yield, height, and weight etc. are the examples of numerical characteristic.

Classification according to Class-Interval. The classification of the given data according to a numerical characteristic is known as the classification according to the class-interval.

In the classification according to class-interval, we must be ascertained before hand whether the data is discontinuous or a continuous one. The discrete variables are easier to deal with. An example of a discrete variable is the throws of a die.

Example(1): Let a die be thrown 50-times, and noted the number at its upper most face each time. If the number obtained are 1, 6, 5, 5, 4, 3, 1, 2, 3, 4, 2, 5, 6, 1, 2, 2, 2, 5, 1, 6, 6, 4, 6, 4, 3, 1, 6, 3, 1, 4, 3, 5, 6, 1, 3, 6, 4, 1, 1, 4, 3, 5, 3, 1, 2, 4, 4, 4, 2, 5, The variable (X) is the outcome of the number at the upper most face of the die in each throw and it clearly assumes the values 1, 2, 3, 4, 5, & 6.

In order to classify the data, these six values assumed by the variable (X) will be written in a horizontal row or a vertical column under the heading "*Variable Value* (X)". Since each value of X is repeated a number of times; the number of repetition of a variable value is *called its frequency* (f). In classification, these respective frequencies for each value of the variable are written against the values of the variable in the 2nd. column under the heading "*Frequency f*". As in the present example, the variable value '1' is repeated 10 times and variable value '2' is repeated 7 times; so the respective frequencies for X=1 & 2 are 10 & 7.

Frequency Table : The table, showing the different values assumed by the variable and their respective frequencies is called a frequency table.

The frequency table for this given example is shown below:—

Table No. 1

Variable value (X)	Frequency (f)
1	10
2	7
3	8
4	10
5	7
6	8
Total	50

This method is adopted only, when the values taken by the variate are not so many (numerous).

On the other hand, when the number of values taken by the variate is large and the range between the greatest and the smallest value is also large, then the entire range is partitioned into classes of appropriate size by drawing the arbitrary lines showing the corresponding frequencies of the variate in each class. Each class is defined by two boundaries, the lower boundary and the upper boundary. The lower boundary of the class is called the *Lower Limit* of the class and the upper boundary is called the *Upper Limit* of the class.

Class-Interval: It is the difference between the upper & the lower limit of the class. For example, for the class 25—35, the upper limit is 35 and lower limit is 25 and the C. I. $=35-25=10$.

Variate Value : The mid-value of the class is called the variate value i.e. $M.V. = \frac{\text{upper Limit} + \text{lower limit}}{2}$ and so in the above mentioned class the $M.V. = \frac{35+25}{2} = 30$

Important points for frequency distribution : While making a frequency distribution, the following points should be borne in mind—

(1) The class-intervals should be equal in each class to provide with easy computations.

(2) The number of classes should not ordinarily exceed 20 and it should not be less than 6. Because for more than 20 classes, the calculations become heavy and there is **not** much gain in the accuracy. Also, for less than 6 classes, there may be a great deal of loss in accuracy.

(3) In general, there should be no class without definite limits as under 'a' or over 'b'.

(4) We must treat all the values assigned to the different classes equal to the mid-value of the classes. Hence the class-interval must be chosen in such a way that the average of all the items in that class should not deviate much from the mid value of the class.

(5) A decision should be made before hand whether the class'a–b'contains a lower limit of the class or b, upper limit of the class. A note of this should also be given, below the table.

(6) The class-interval should be an integer as far as possible.

Methods of Classification according to Class-interval:–

There are two methods of classification according to the class-interval.

(1) Exclusive Method (2) Inclusive Method

(1) **Exclusive Method**:—In this type of classification, the upper limit of any class is the lower limit of the succeeding class i.e. the class-limits are of the type a–b, b–c, c–d. etc. Hence, there is a confusion about those items which are exactly equal to the class-limits (called border line items). Usually, the border line items are placed

in the classes where they are as the lower limits of the class. As variable value 'b' is placed in the class b–c. This confusion can be removed by having the classes of the type, a and under b, b and under c and so on.

(2) **Iuclusive Method**:—In this type of classification there is a gap between the upper limit of the class and the lower limit of the succeeding class. Hence, it leaves no room for confusion for the border line items. Thus the classification will be of the form a–b, c–d, e–f etc. and the class a–b will involve all the items from a to b inclusive a & b within it. This method can not be used in the case of a continuous variate.

Rules for making a frequency distribution:

(1) Before starting the actual work of classification, a preliminary inspection of the data should be made and the difference of the highest & the lowest values should be divided by the desired number of the class to be formed. It will give an approximate idea of the class-interval and will also help in setting the class limit.

(2) After deciding about the class-interval and class-limits, a table containing three headings namely, class-interval, tally marks & frequency should be prepared.

(3) Read for the items in the row table and for each items put the mark '1' in its corresponding class in the column headed *Tally marks*. After putting 4 marks of the above type the fifth is obtained by crossing the 4 marks and thus forming a group of 5 which helps in counting in the end.

(4) The sum of the tally marks is written in the column of frequency against the respective classes.

(5) The check of the sum of all the frequencies (Σf) must show Σf equal to the total number of variate-values.

Example (2)

The marks obtained by 60 students in the paper of statistics are given below—

22	47	9	42	31	17	13	15	18	25	35	33
38	15	0	33	10	34	29	26	16	33	16	27
24	22	26	19	14	36	18	25	21	12	21	10
28	25	17	38	10	3	31	24	3	2	36	18
26	29	27	39	28	35	26	27	13	33	25	18

Arrange the data in the form of a frequency distribution?

Sol:—

Here the variate is the numbers obtained in statistics and the original form of the data is an *ungrouped data*. If we arrange the data in the form of a table shown in the example (1), then there are as many as 29 values of the variable. Some of them occuring once, some twice, some thrice and some 4 times at the most.

Though, it is a frequency distribution giving a clear idea when compared to an ungrouped data, yet it may be improved by classifying it into groups. From an eye inspection of the data it is obvious that the range is 47–0=47, because the highest value is 47 and the lowest value is 0. Further, the variable does not assume a fractional value and is a discrete variate having gaps between its successive values. The lowest value 0 suggests that we must start with lower limit as Zero and continue by 4 marks interval, adopting the Inclusive-Method of classification.

The frequency-table will be formed as given below—

Table No. 2

Class-interval	Tally marks	Frequency (f)
0—4	𝍷𝍷𝍷𝍷	4
5—9	𝍷	1
10—14	𝍸 𝍷𝍷	7
15—19	𝍸 𝍸 𝍷	11
20—24	𝍸 𝍷	6
25—29	𝍸 𝍸 𝍸 𝍷	16
30—34	𝍸 𝍷𝍷	7
35—39	𝍸 𝍷	6
40—44	𝍷	1
45—49	𝍷	1
Totals		$\Sigma f=60$

In a similar manner, we can form the frequency table (distribution) when the ungrouped data is such that it pertains to a continuous variate and there *Exclusive Method* of classification will be used.

Example (3)

The following table gives the Vickers Hardness number of 20 shell cases—
66.3 61·3 62·7 60·4 60·2
64.5 66·5 62·9 61·5 67·8
65.0 62·7 62·2 64.8 65·8
62·2 67·5 67·5 60·9 63·8. Arrange the data in the form of a frequency distribution ?

Solution:—

The largest value is 67·8 and the lowest is 60·2, so the range is 67·8—60·2=7·6; it suggests us to start with 60·0 as our lower limit and to continue by **1** (unit) class interval. The variable is the Hardness Number and clearly it is a continuous variate.

The frequency- table is given below—

Table No 3

Class-interval	Tally marks	Frequency (f)
60—61	|||	3
61—62	||	2
62—63	𝍸	5
63- 64	|	1
64—65	||	2
65—66	||	2
66—67	||	2
67—68	|||	3
Totals	—	$20 = \Sigma f$

Cumulative Frequency Table: In some of the statistical investigations (Educational tests, wages or salary statistics etc.), we require the number of variates which are *'less than'* or *'more than'* a given value. For this purpose, it is necessary to change an ordinary frequency table into a cumulative frequency table. It can be done in the following two ways:—

(1) Less than type:

Suppose, we are given the daily expenses for one week as follows:—

Table No. 4

Days	Expenses (Rs.)
Monday	4
Tuesday	6
Wednesday	10
Thursday	16
Friday	12
Saturday	8
Sunday	4

Now if we want to know the total expenses upto Wednesday then it will be given by the sum of the expenses for Monday, Tuesday & Wednesday i.e. $4+6+10=20$. Similarly the total expenses incurred upto any day of the week can be obtained by adding the expenses from Monday upto that day. In the tabular form, the total expenses can be represented as given below in table No. 5.

Table No. 5

Days	Expenses up to the day (Rs.)
Monday	4
Tuesday	10
Wednesday	20
Thursday	36
Friday	48
Saturday	56
Sunday	60

On the other hand, the expenses after any day (including the day) will be obtained by adding the expenses for that day and after-wards. The following table No. 6 serves this purpose.

Table No. 6

Days	Expenses after the day (Rs.)
Monday	60
Tuesday	56
Wednesday	50
Thursday	40
Friday	24
Saturday	12
Sunday	4

It the days are replaced by the variate x, by its values $x_1, x_2 \ldots\ldots\ldots x_7$. then the values in the 2nd. column of table No. 5 show the number *'less than'* the variate value and column 2nd. of the table No. 6 shows the number *'more than'* the variate value.

These frequencies, which are less than or more than the variate value (as shown in second column of table No. 5 & 6) are called the *'Cumulative Frequencies.*

The similar procedure is adopted when instead of the variate value, the classes are given.

Example No. 4

Form a cumulative frequency table from the following data:—

class:	0–3	3–6	6–10	10–12	12–15	15–19	19–20	20–23	23–25
Frequency:	2	4	5	7	8	11	12	14	10

Solution:—

(i) In making the cumulative frequency table of "*Less than type*", the first class will be replaced by under 3, second class by under 6 and so on. The cumulative frequencies will be the sums of the frequencies upto that class for all the respective classes. The cumulative frequency tabale of "*Less than type*" is given below:—

Table No. 7

Variate value	Cumulative frequency (c. f.)
Under 3	2
Under 6	6
Under 10	11
Under 12	18
Under 15	26
Under 19	37
Under 20	49
Under 23	63
Under 25	73

(2) **More than type:**

(ii) In foming the cumulative frequency table of "*more than type*", the first class is replaced by more than 0, second by more than 3 and so on. The cumulative frequencies are the sums of frequencies onwards this class for all the respective classes. The cumulative frequency table of "*more than type*" is given below:—

Table No. 8

Variate Value	Cumulative frequency (c. f.)
More than 0	73
More than 3	71
More than 6	67
More than 10	62
More than 12	55
More than 15	47
More than 19	36
More than 20	24
More than 23	10

Note: (1)—

From the above example, it is clear that in the "Less than type" c. freq. table, we have to replace the classes by 'under' or 'below' the 'upper limits of the classes' and in the "More than type" c. freq. table we have to replace the classes by 'More than or over' the lower limits of the *c lasses'.*

Forming an ordinary freq. table from a c. freq. table:—

There are the cases when we want to change the cumulative frequency table into an ordinary frequency table. For example, if for every unit of consumption of a certain commodity, its total utility is given and from it we want to know the marginal utility for every unit of consumption then it will be a case of forming an ordinary freq. table from a cumulative one.

Example (5)

In the table given below, the total utility for a unit consumption of a certain commodity is given. Find the marginal utility for every unit of consumption ?

Unit :	1	2	3	4	5	6	7	8	9	10	11	12
Total utility :	3	11	21	33	49	63	73	81	98	103	107	108

Solution:—

We know that:—

the total utility for ith unit = the marginal utility for Ist unit + ... + the marginal utility for i^{th} unit.

From this, it is clear that the relation between the total utility and the marginal utility is the same as between the cumulative frequency and the ordinary frequencies.

Here we have to convert a cumulative frequency table into an ordinary frequency table as shown below:—

Table No. 9

Unit	Marginal utility	Unit	Marginal utility
1	3	7	10
2	8	8	8
3	10	9	17
4	12	10	5
5	16	11	4
6	14	12	1

Note: (2)—

To form the ordinary frequency table from cumulative frequency table of "less than type" we have—

frequency for i^{th} class = c. f for i^{th}. class — c. f. for $(i-1)^{th}$ class.

Similarly, to convert a c. f. table of "more than type' into an ordinary ferquency table we have

frequency for i^{th} class = c. f. for $(i)^{th}$ class c. f. for $(i+1)^{th}$ class.

Example (6)

From the following data, find the number of persons in the clas es 20--25, 25--30...... ..and 55 -60.

Age more than :	20	25	30	35	40	45	50	55
No. of persons :	800	750	680	580	400	250	130	60

Solution:—

We have been given that the number of persons who are old than 20 is 800 and those who are old than 25 is 750. Thus the no. of persons who are between the age of 20 & 25 is 800–750 = 50 only. In the same way, the no. of persons between 25–30, 30–35 .... and 55–60 can be found. Because the formula used is:—

frequency for i^{th} class = c. f. for i^{th} class — c. f. for $(i+1)^{th}$ class.

Thus the table on the next page shows the no. of persons between the different age-groups.

Tablo No. 10

Age-group	No. of persons
20—25	50
25—30	70
30—35	100
35—40	180
40—45	150
45—50	120
50--55	70
55—60	60
Total	800

Example (7)

Obtain the ordinary frequency table from each of the following tables:—

(1)

	No. of Students
Marks Below 10	3
,, ,, 20	8
,, ,, 30	17
,, ,, 40	20
,, ,, 50	22

(2)

	No. of Students
Marks above 0	30
,, ,, 10	26
,, ,, 20	21
,, ,, 30	14
,, ,, 40	10
,, ,, 50	0

Solution:—

(1) Here the c. f. table of 'Less than type' is given and we have to convert it into an ordinary frequency table. The frequency for i^{th} class=c. f. for i^{th} class—c. f. for $(i-1)^{th}$ class. Thus, the table given below shows the desired no. of students in the different groups of marks:—

Table No. 11

rks	No. of students
0—10	3
10—20	5
20—30	9
30—40	3
40—50	2
Total	22

(2) Here the c. f table of "More than type" is given and we have to convert it into an ordinary frequency table. The frequency for i^{th} class=c. f. for i^{th} class—c. f. for $(i+1)^{th}$ class. Thus, the table given below shows the desired no. of students in the different groups of marks:—

Table No. 12

Marks	No. of students
0—10	4
10—20	5
20—30	7
30—40	4
40—50	10
Total	30

Tabulation:—

It has been mentioned already that the classification alone is not sufficient for comparisons and inferences. The data is further put to the treatment of tabulation which make it fit for further analysis and comparisons.

Thus the *Tabulation* is the process of arranging the classified data in an orderly manner into rows & columns in such a way that it makes clear all the important informations contained in the data.

For tabulation, the following points should be kept in mind—

(1) Every table should have a title, explainatory in itself and it should provide the informations regarding:—

(a) What the data are. (b) Where the data are.

(c) Principle of clasification. (d) Time of data.

(2) The rows & columns must be arranged in a logical order to facilitate the comparisons.

The headings & sub-heading should be concise and without any ambiguity.

(4) The units of the data presented, should be clear.

(5) The complicated tables should be avoided and so they must be broken up into parts, in more than one simple tables.

(6) The size of the table shonld suit the size of the paper.

(7) The table should be presented in a neat, clean and comprehensive way by drawing the double lines wherever required.

(8) The source of the table should also be written.

Types of Tabulation

There are two types of tabulation:—

(1) Simple tabulation (2) Complex tabulation.

Simple Tabulation:—

In a simple tabulation, the data is presented with respect to one character only or in other words, a simple table is capable to provide the information relating to one character only. For example, the population of India can be divided according to religion. The table for this purpose will be as follows:—

Table No. 13

Religion	Population
1. Hindu	...
2. Muslim	....
3. Sikh	...
4, Christian	...
5. Jain	...
6. Others	...
Total	...

(2) **Complex Tabulation:—**

A complex table gives the information relating to several characters. This type of tabulation is further classified as 'Double Tabulation' 'Triple or Treble Tabulation' and 'Manifold Tabulation'.

Double or two fold Tabulation:

In this type of tabulation, the data is divided according to two characteristics. For example, the population of India can be divided according to religion & States and the table for it is given below:—

Table No. 14

States	Population						Total
	Hindu	Muslim	Sikh	Christian	Jain	others	
1. Assam							
2. Bihar							
3. Punjab							
4. U. P.							
....							
....							
Total							

Treble Tabulation :

In this type of tabulation, the data is sub-divided according to three characteristics. Treble tabulation is capable of answering three mutually dependent qnestions. The table given below is an example of treble tabulation. It shows the distribution of India's population according to States, religion and sex.

Table No. 15

States	Population														
	Hindu			Muslim			Sikh			Others			Totals		
	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
1. Assam															
2. Bihar															
3. Punjab															
4. U P.															
....															
....															
....															
....															
Total															

Manifold Tabulation :

In this type of tabulation. the data is divided with respect to more than three characteristics. From this type of table, we can get the information relating to several (more than three) characteristics. An example of this type of tabulation is the division of India's population according to States, religion, sex, literacy and age etc.

Example (8)

Draw up in detail, with proper attention to spacing, double lines etc. and showing all the sub-totals, a blank table, in which could be entered the numbers occupied in the six industries at two dates (July 63, July 64) distinguishing males from females and among the latter,single, married and vidowed.

Solution:—

The desired table is given below:—

Table No. 16

Industry	No. in July,63					No. in Jnly,64					Totals					Totals
	Male	Female			Total	Male	Female			Total	Male	Female			Total	
		Single	Married	Vidowed			Single	Married	Vidowed			Single	Married	Vidowed		
1.																
2.																
3.																
4.																
5.																
6.																
Total																

The above table shows the no. of employees in six industries according to their distribution into sex and two dates.

Example (9)

Draw up two independent blank tables, giving rows, columns, and totals in each case, summarizing the details about the number of families, distinguishing males from females, earners from dependents and adults from the children.

Solution:—

The desired table showing all the required informations will be a complex table and it will be tedious to understand it. For the sake of convenience we partition this table into two simple tables:—

(i) Giving the distribution of all the male members in the families according to age and earning;

(ii) giving the distribution of all the female members in the families according to age and earning.

(i) **Table**—Showing the distribution of male members according to age and earning—

Table No. 17 (i)

Family	Earner			Dependent			Totals		
	Adults	children	Total	Adults	children	Total	Adults	children	Total
1.									
2.									
3.									
4.									
5.									
6.									
......									
....									
Totais									

(ii) **Table:**—Showing the distribution of female members according to age and earning—

Table No. 17 (ii)

Family	Earner			Dependent			Totals		
	Adults	children	Total	Adults	children	Total	Adults	children	Total
1.									
2.									
3.									
4.									
5.									
......									
......									
Totals									

Chapter II

Graphs And Diagrams

After condensing and summarizing the complex and numerous data in a systematic manner by means of classification and tabulation, we require certain divices which may present the condensed form of the data. This condensed form of the data is such that it becomes at once comparable and leaves an everlasting impression on the brain of the observer. One of the methods making the data intelligible is to represent it by means of graphs and diagrams. The graphic & diagrammatic representation of the data is always appealing to the eye as well as to the mind of the observer.

Advantages of Graphs & Diagrams:—

(1) The graphs & diagrams represent the data in attractive and appealing way both to the eye and mind.

(2) The graphic & diagrammatic representation of the data leaves an everlasting effect on the brain.

(3) The diagrams are not only attractive and impressive but they save time also. Because through the diagrams, it is possible to have an immediate grasp of significance.

(4) Another merit of the diagrams is the ease with which the two sets of the data are compared with each other.

(5) Forecasting becomes easier with the help of the graphs.

(6) The graphs are helpful in the interpolation also and they give an indication of correlation between the two variables.

(7) The partitioning values (median & quartiles etc.)

and mode are also determined by graphs.

Limitations of graphs & diagrams:—

(1) The graphs and diagrams provide an approximate picture of the data and not the accurate one. Thus they are useful in representation to general public but not to a statistician or an investigator who is interested in the detailed study.

(2) They are used to compare only such data which are technically comparable.

(3) They are easily capable of misuse.

Graphical Representation of Frequency Distributions:

In the graphical representation of the frequency distributions the horizontal axis of the graph is used to show the variate values and the vertical axis for the frequencies of the variate. The graphs of the frequency distributions are of the following types—

(1) Histogram (2) Frequency-polygon and Frequency Curve (3) Cumulative frequency curve or ogive.

(1) **Histogram** : It is composed of a set of rectangles one over each class-interval on the X-axis. The width of the rectangle is taken proportional to the class-interval and its height is taken proportional to the frequencies in the case of equal intervals. But in the case of unequal class-intervals the height of a rectangle is taken proportional to the ratio of frequencies to the class-interval and thus the area of the rectangle is proportional to the frequencies of the variate or the classes.

Example (1) :—

The following is the frequency distribution of the yield of sugar-cane in tons per acre.

Class :	35-40	40-45	45-50	50-55	55-60	60-65	65-70
Frequency :	7	8	12	26	32	15	9

Drawn a histogram representing the above distribution ?

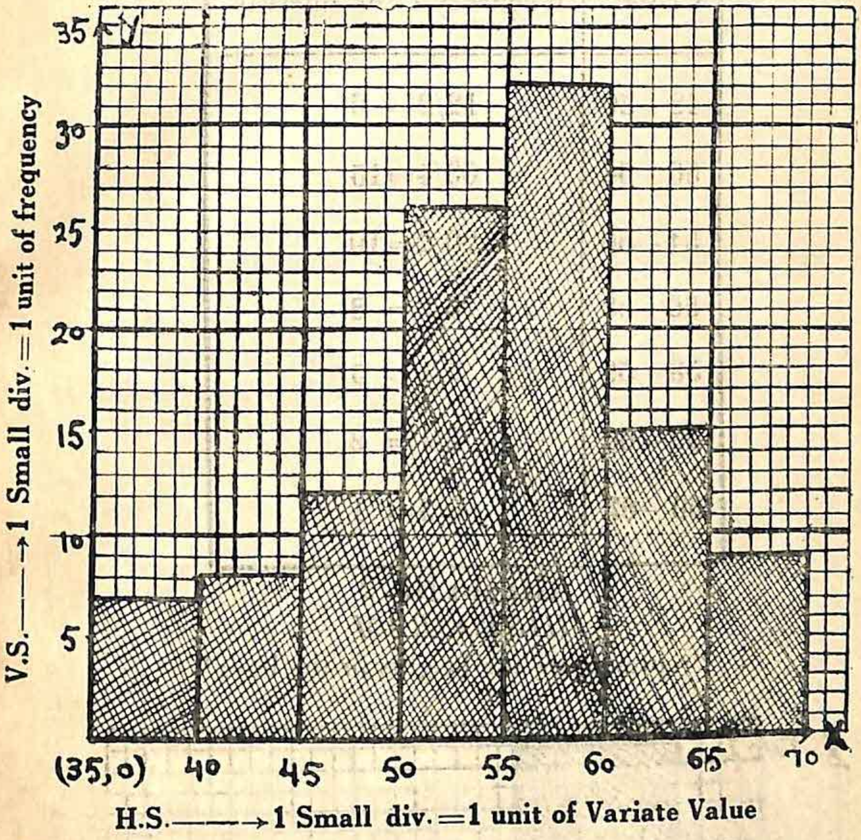
Solution:—

In this example, the class intervals are equal and so the heights of the rectangles will be proportional to the frequencies.

Graph No. 1

HISTOGRAM

"representing the yield of sugarcane in tons/acre"



Example No. (2): –

Draw the histogram of the following distribution:—

Heights : (in inches)	48–50	50–54	54–56	56–58	58–59	59–60	60–63
No. of Boys : (frequency) :	12	60	20	16	5	2	3

Solution :—

In this example, the class-intervals are not equal, so the heights of the rectangles will be proportional to the ratio of the frequencies to their respective class-intervals. Thus the heights of the rectangles are given in the following table against the respective classes:—

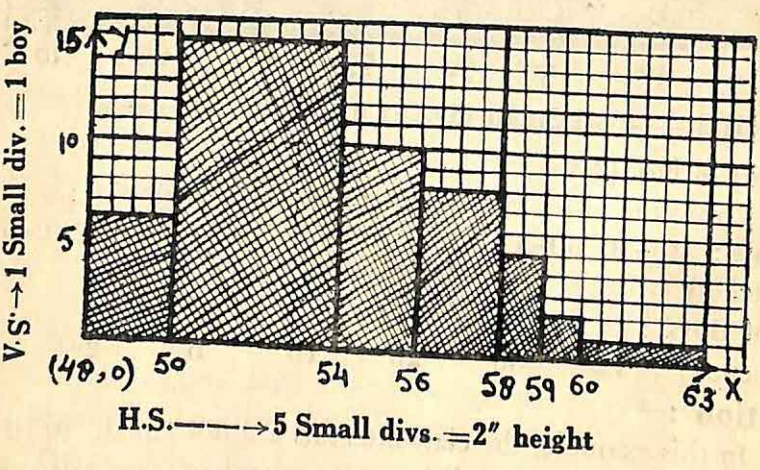
Table No. 18

Classes	Frequency/Class Interval
48—50	12/2= 6
50 - 54	60/4=15
54—56	20/2=10
56—58	16/2= 8
58—59	5/1= 5
59—60	2/1= 2
60—63	3/3= 1

Graphs No. 2

HISTOGRAM

"representing the heights of the boys"



H.S.———→5 Small divs.=2″ height

Frequency Polygon & Frequency Curve:—The frequency polygon is constructed by joining the points by means of straight lines whose abscissae are the mid points of the classes and the ordinates are the corresponding frequencies. A frequency polygon can also be obtained by joining the mid-points of the upper sides of a histogram by straight lines.

A frequency curve is a smooth, free hand curve, drawn through all the points which are obtained for the frequency polygon. The area under the curve should be equal to that of the histogram. Frequency polygon is used to find the 'mode' which is the *apex* of the curve.

Example (3a)—

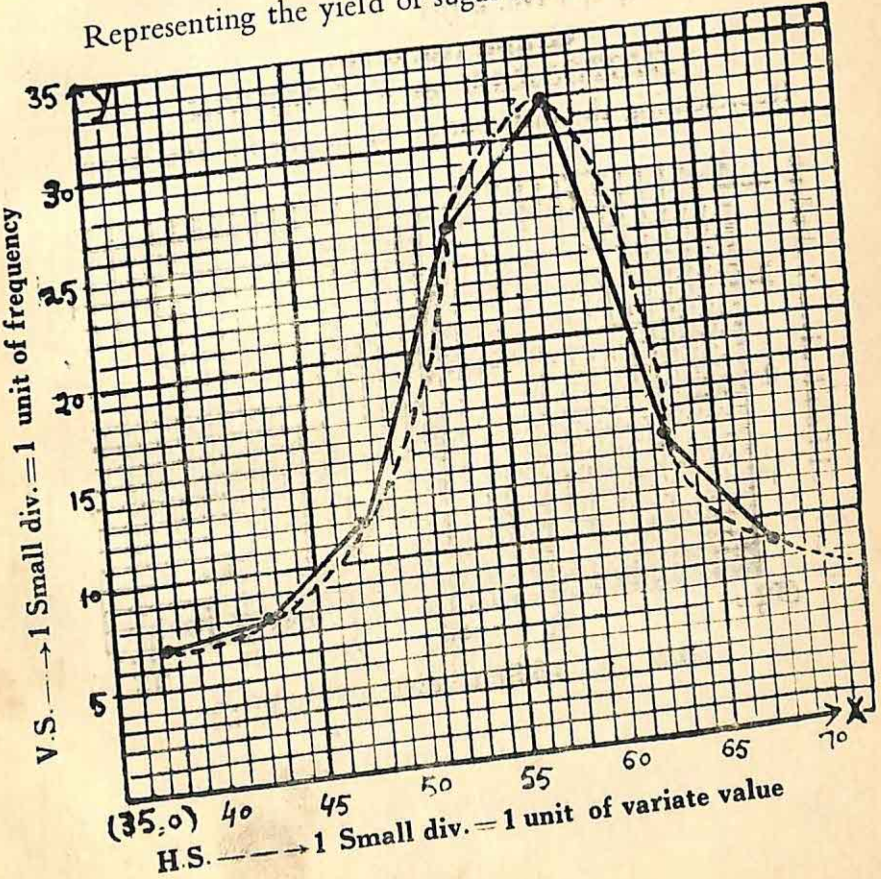
Draw a frequency ploygon & a frequency curve for the data in example (1) on page 28.

Solution:—

In this example, the points of the classes are plotted on the X-axis and the frequencies on the Y-axis.

Graph No. 3 **(a)**

Representing the yield of sugar cane in tons/acre.



Example 3 (**b**):—

The following table gives the distribution of height in inches for 100 students.

Interval	Frequency
>57 upto 60	3
>60 ,, 63	12
>63 ,, 66	31
>66 ,, 69	37
>69 ,, 72	16
>72 ,, 75	1
Total	100

Represent the data in the form of a histogram as well as a frequency polygon ?

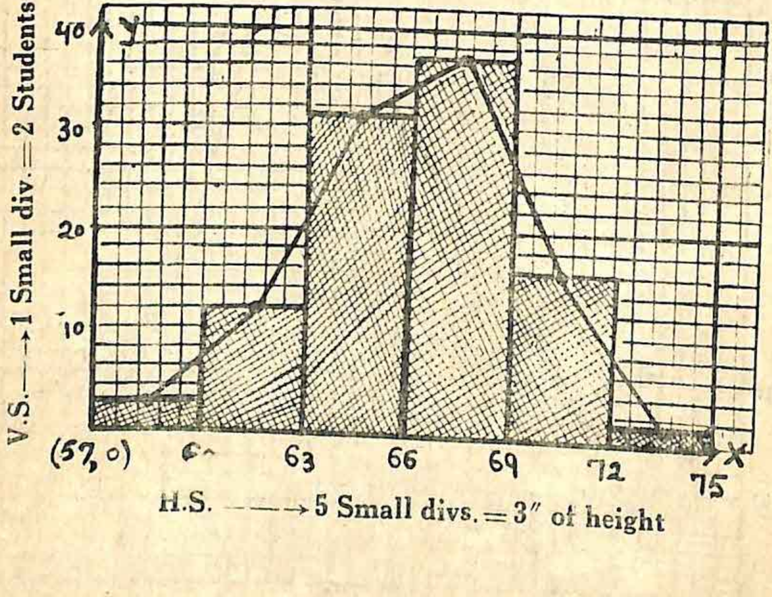
Solution:—

See the graph No. 3 (b)

Graph No. 3 (**b**)

HISTOGRAM & F. POLYGON.

Representing the distribution of height for 100 students



(3) **Cumulative frequency curve or ogive** : The ogive is constructed from a less-than type cumulative frequency table. The upper limits of the classes are taken as abscissae and the corresponding cumulative frequencies as ordinates and thus all the points are plotted on the graph. The free hand smooth curve through all these points is called the *ogive*. Ogive is helpful in determining the partitioning values (median, quartiles etc.).

Example (4):—

Draw a cumulative frequency curve for the data given in example (1) on page 28.

Solution:—

First of all, we shall construct the cumulative frequency table from the given data as follow:—

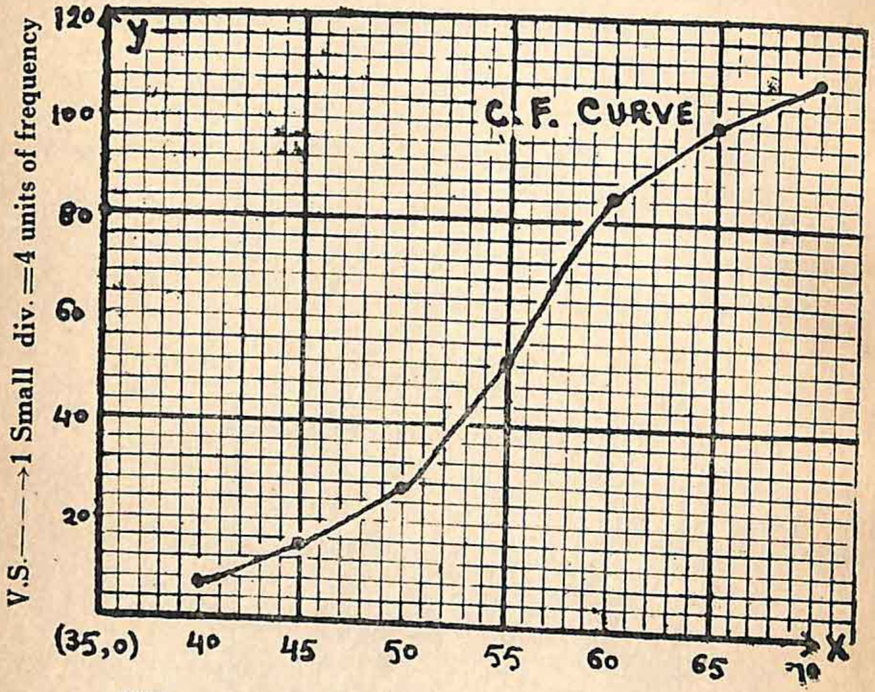
Table No. 19

Variate Value Under	Cumulative frequency(c.f.)
Under 40	7
Under 45	15
Under 50	27
Under 55	53
Under 60	85
Under 65	100
Under 70	109

The cumulative frequency curve is shown below:—

Graphs No. 4

Representing the yield of sugar-cane in tnos/acre



H.S.——→1 Small div. =1 unit of variate value

Gap between the upper & the lower limits of the adjacent classes:—

In the case of the grouped data of the type a-b, b-c, c-d, and so on, we have seen that the graphs of the above three types can easily be used. But in case, where there is a gap between the upper and the lower limits of the adjacent classes, the class limits should be so modified that the gap vanishes and the upper & lower limits of the adjacent classes become the same.

Example (5):—

Draw the histogram, frequency Polygon and the ogive for the following data:—

Scores :	20-29	30-39	40-49	50-59	60-69	70-79
No. of students :	2	14	22	20	14	3

Solution:—

In this example, there is a gap between the upper & the lower limits of the adjacent classes. Hence, first we modify the classes to have the continuous from of the data and thus finish the gap. For the construction of histogram, frequency polygon & the cumulative frequency curve for this data, we form the following table:—

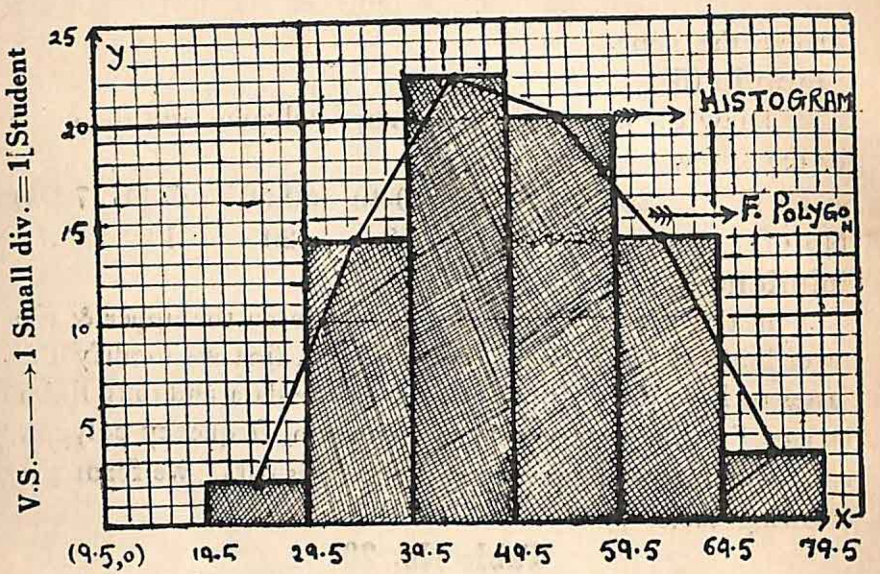
Table No. 20

Classes	Frequency (f)	Comulative frequency (c.f.)
19·5—29·5	2	2
29·5—39 5	14	16
39·5—49·5	22	38
49·5--59·5	20	58
59·5—69·5	14	72
69·5—79·5	3	75

Graph No. 5 (a)

HISTOGRAM & F. POLYGON

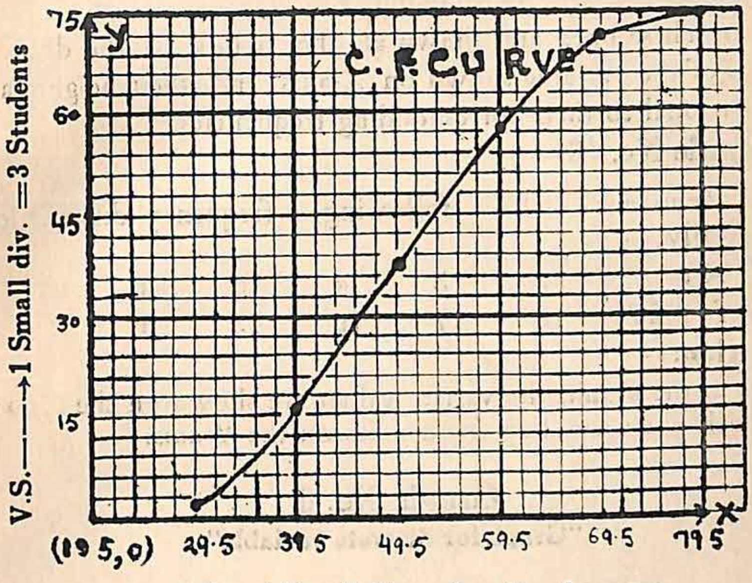
"representing the scores for Students"



H.S.——→1 Small div.=2 units of scores

Graph No: 5 (b)

Cumulative-Frequency-Curve
Or
OGIVE



Graph For Discrete Variable:—When the data is given for a discontinuous variable accompanied by its corresponding frequencies, then the graphical representation of such a data is made by drawing thick lines parallel to the Y-axis. These lines are drawn at the points of the discrete values of the variable shown on X-axis and their heights are proportional to their corresponding frequencies.

Example No. (6)

Represent the following frequency-distribution graphically:—

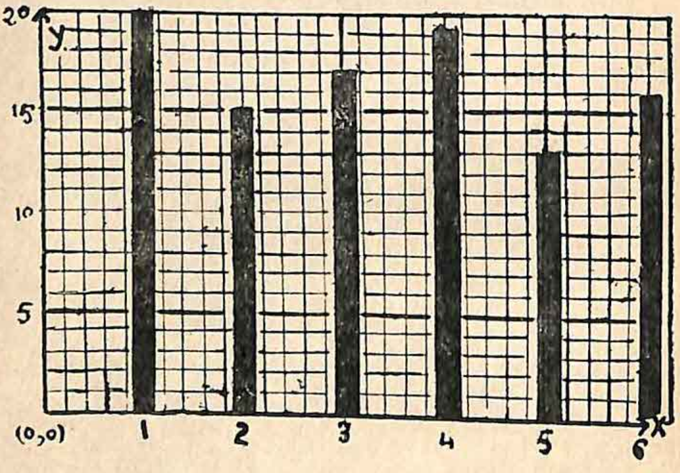
X :	1	2	3	4	5	6
f :	20	15	17	19	13	16

Solution:—

In the graph, the variate values are shown on the X axis and their corresponding frequencies on the Y-axis.

Graph No. 6

"Graph for discrete variable"



V.S.→1 Small div. = 1 unit of frequency

H.S.———→5 Small divs. = 1 unit of variate value

Graph for Time Series : (Historigram) The Historigram is the graph obtained from the time series data. In these graphs, the time variable 't' is measured along X-axis and its corresponding values 'u_t' along Y-axis. All the available points form the data are plotted on the graph and joined by means of straight lines.

Following few examples will illustrate the type of graph:—

Example No. (7): –

The following table provides the yearly figures of production of sugar in lakh tons—

Year t	production of sugar (lakh tons) u_t
1948	9·65
1949	8·90
1950	6·75
1951	10·50
1052	11·25
1953	10·70
1954	12·35
1955	12·55
1956	12·95
1957	13·00
1958	13·50

(i) Draw the graph to show the fluctuations of the production of sugar.

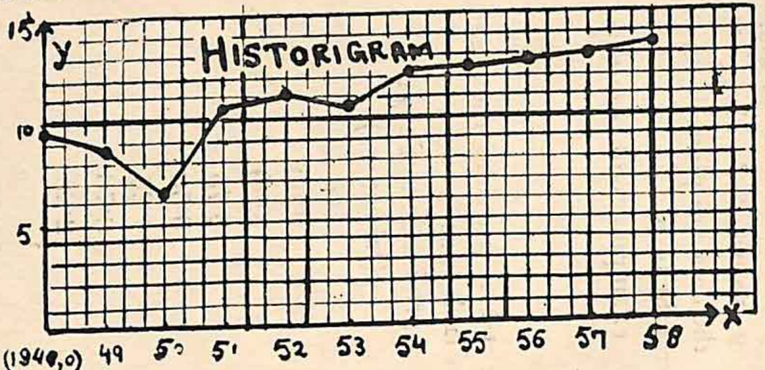
(ii) Comment on the fluctuations of the production.

Solution:—

(i) The production of the sugar in lakh tons is taken along Y-axis and years on the X-axis in the following graph.

Graph No. (7)

Representing the production of sugar in lakh tons, from 1848—58.



V.S.——→1" = 10 lakh tons

H.S.——→3 Small divs. = 1 year

Example No (8):—

Represent the figures given below, on a graph-paper and comment on trend shown by the data.

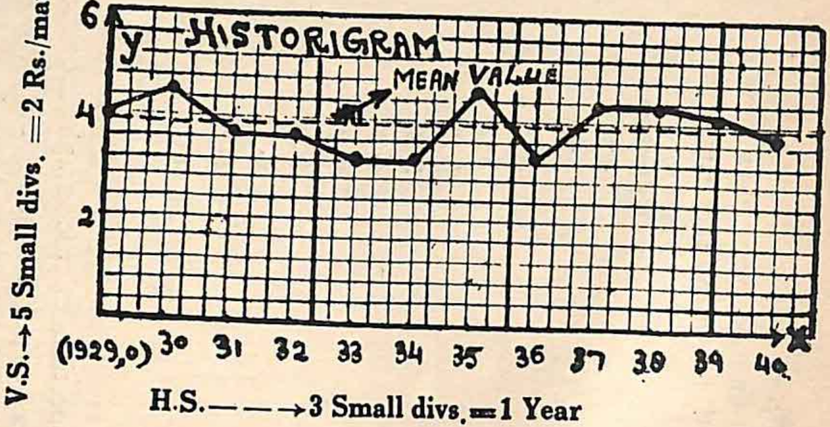
Year	Price for Arhar in Rs./maund.
1929	4·0
1930	4·6
1131	3·6
1932	3·6
1933	3·3
1934	3·3
1935	4·7
1936	3·4
1937	4·3
1938	4·3
1939	4·2
1940	3·9

Solution :

An eye inspection of the graph shows that the prices are neither decreasing nor increasing but they are fluctuating around the price Rs. 3.9/maund, the mean value.

Graph No. (8)

Representing the price for Arhar in Rs./maund from 1929 to 1940



Example No. (9):—

Draw the graph from the following table, in which the growth of the three kinds of wheat may be read from the graph. Discuss the growth of the three kinds of wheat ?

(U. P. Board 1956)

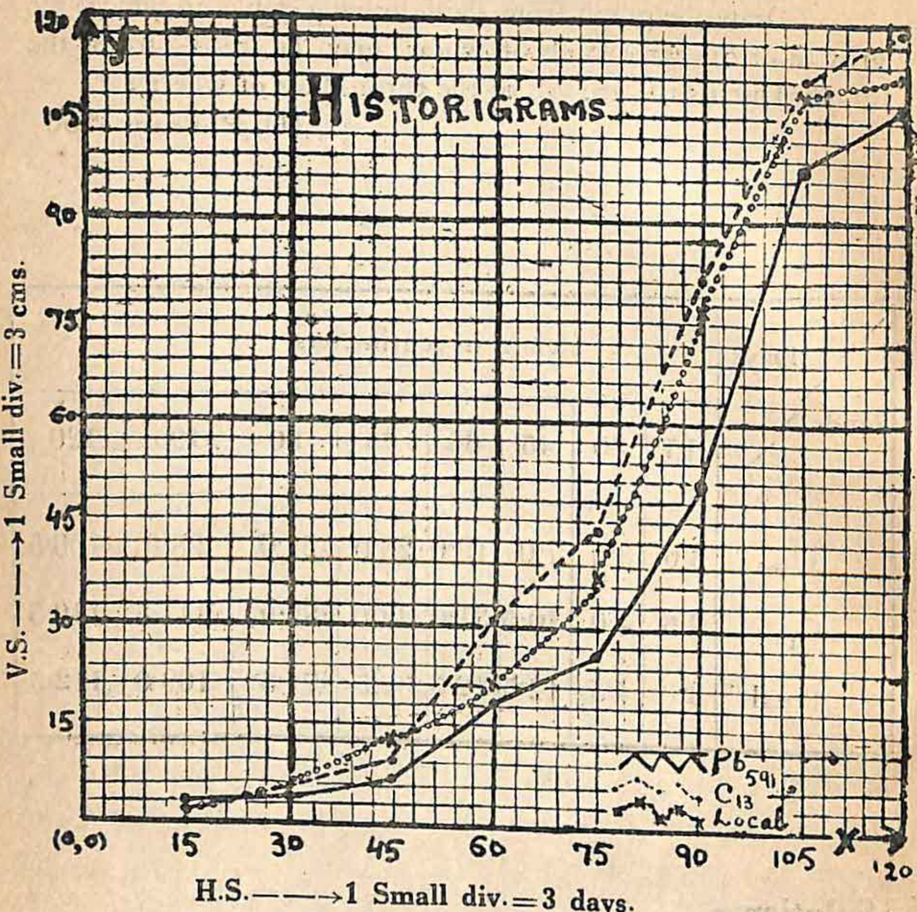
Age Days Kinds	Height in centimeters							
	15	30	45	60	75	90	105	120
Pb_{591}	3·0	3·5	7·0	19·0	26·0	51·0	98·0	106·5
C_{13}	2·5	4·0	10·0	31·0	44·0	82·0	111·5	119·5
Local	3·0	4·5	12·5	22·0	37.5	77·5	109·0	112·5

Solution:—

Of all the three varieties of wheat, the growth in C_{13} is minimum upto the age of 45 days in comparison to the local variety and after this, the growth of C_{13} is the highest, while that of Pb_{591} is minimum. The growth of the local variety is between the two.

Graph . (9)

Representing the growth of three kinds of wheat



An inspection of the graph clearly shows that upto the age of 45 days, the growth in local variety is the highest and after it that of C_{13} is the highest. Thus, as regards the character of growth, C_{13} is the best.

Example No. (10):—

The average yields in maunds per acre of rice and wheat crops in a state of India since 1950-51 are given below. Comment whether the yields are increasing or decreasing ?

Year	Rice	Wheat
1950—51	5·27	8·88
1951—52	4·38	8·25
1952—53	5·31	9·21
1953—54	6·46	9·15
1954—55	6·12	9·64
1955—56	7·18	8·26
1956—57	6·21	8·46
1957 - 58	6·19	7·89
1958—59	7·13	8·61
1959—60	6·15	9·21
1960—61	7·15	10·91

(M.Sc. Ag. 1962)

Solution:—

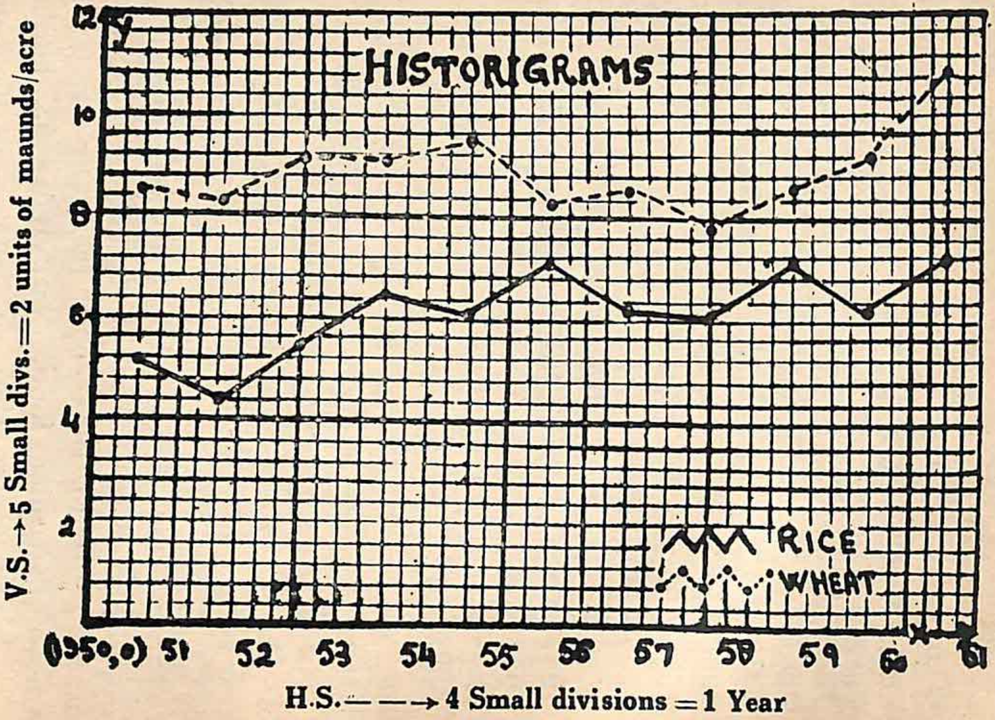
Let us first plot the yields of rice and wheat since 1950—51 to 1960—61 on the graph-paper. As a result of this, we obtain the following graphs.

An eye-inspection of the above graphs tell us that—
(i) the yield of wheat has an increasing tendency. The yield from 1950—51 to 1954—55 is increasing except a slight downward fluctuation for the year 1951—52. Further the yield is decreasing upto 1957—58 and again increasing for the remaining years.

(ii) The yield of rice is increasing from 1950—51 to 1955—56 except a slight downward fluctuation for the year 1954—55. For the next two years, the yields are decreasing and are the same. For the year 1958—59, the yield increases and again decreases for the year 1959—60 and increases for 1960—61. Thus on the whole, we can say that the yield shows an increasing trend.

Graph No. (10)

Representing the yield of Rice & wheat from (1950-51) to (60-61)



Example No. (11):—

The following table provides the yearly figures of production in lakh tons in U. P. Comment on the wheat production in U. P.

Year	Yield	Year	Yield
1943—44	19.01	1952—53	28.24
1944—45	23·56	1953—54	31·06
1945—46	22·83	1954—55	32·84
1946—47	23·37	1955—56	30.41
1947—48	21·94	1956—57	31·15
1948—49	19·74	1957—58	27·06
1949—50	24·26	1958—59	30·36
1950—51	26·78	1959—60	32·42
1951—52	25·33	1960—61	38·82

(M.Sc. Ag. 1963)

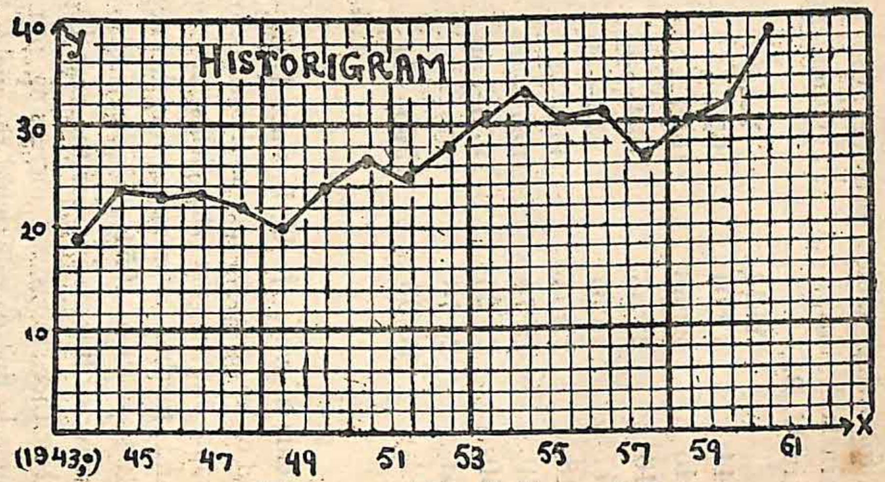
Solution:

The yields of wheat in U. P. since 1943—44 to 1960—61 are plotted on the graph-paper as shown below.

The graph shows that the yields are increasing upto 1945—46 and further decreasing for the next two years. Again from 1948—49 to 1954—55 the yields are increasing except a downward fluctuation for the year 1951—52. After this, the yield neither shows an increasing nor a decreasing tendency, it merely fluctuates around the yield 30 lakh tons. But for the year 1960—61, the yield increases. Thus on the whole, there is an increasing trend in the production of wheat.

Graph No. (11)

Representing the production of wheat in lakh tons from (1943-44) to (1960-61)



H.S.——→4 Small divs. = 2 Years

V.S.→1 Small div. = 2 lakh tons

Diagrams

The following are the important diagrams—

(1) One dimensional diagrams,
(2) Two dimensional diagrams,
(3) Three dimensional diagrams,
(4) Cartograms and
(5) Pictograms.

(1) **One Dimensional Diagrams**:—The one dimensional diagrams are the lines or bars arranged in ascending or descending order of magnitudes (length) along a vertical or a horizontal scale. The lengths of the bars are proportional to the magnitudes of the items.

(2) **Two Dimensional Diagrams**:—In the two dimensional diagrams, the magnitudes of the items are represented by the areas as by squares, rectangles and circles etc

(3) **Three Dimensional Diagrams**:—In the three dimensional diagrams, the data are represented by the cubes or cylinders and the magnitudes of the items are represented by the volumes of the cuboids or cylinders.

(4) **Cartograms**:—Here the data is presented by means of maps.

(5) **Pictograms**:—Here the data is presented by means of pictures.

But the bars and circular diagrams are most commonly used because of their accuracy and easiness in sketching.

"One Dimensional Diagrams"

The bar diagrams of common use in the one dimensional diagrams are of the following main types:—

(1) Simple bar diagrams,
(2) Multiple bar diagrams,
(3) Sub-divided bar diagrams.

(1) **Simple bar diagrams**:—Here the magnitudes of the items are represented by thick bars of uniform width with equal spacing between any two consecutive bars. The lengths of the bars are proportional to the magnitudes of the items.

One bar is drawn for each item and they are arranged in the ascending or descending order of lengths along a vertical or horizontal line. These simple bar diagrams are appropriate for the data contained in simple tables.

Example (12)

Draw a simple diagram to represent the following statistics relating to the area under different crops:—

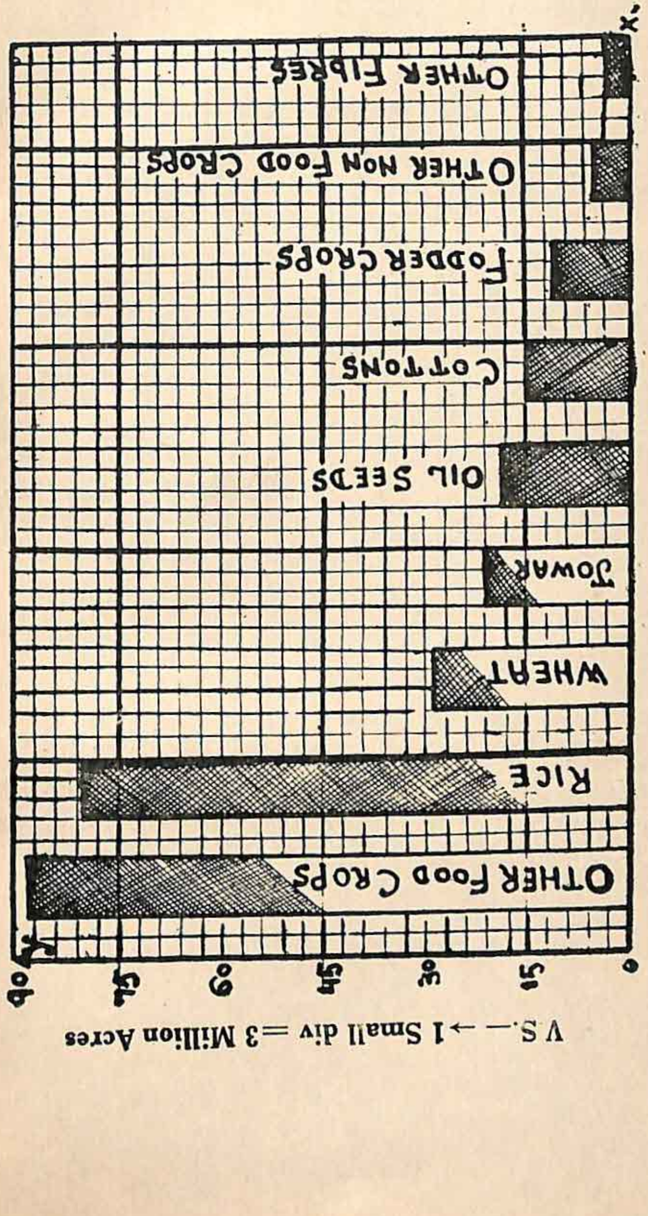
Crops.	Million Acres
Rice	80·3
Wheat	27·6
Jowar	21·4
Other food crops	88·2
Oil seeds	17·6
Cotton	14·5
Other fibres	3·1
Fodder crops	10·2
Other non food crops	3·9

Solution:—

The Acreage (million acres) under different crops is shown by the simple bar diagrams:—

Diagram No. 1

"Simple Bar Diagrams for Acreage under different Corps"



H.S.——→3 Small divs.=width of each bar and 2 Small divs.=Spacing between two consecutive bars.

V.S.——→1 Small div =3 Million Acres

(2) **Multiple Bar Diagrams:**—These bar diagrams are constructed like simple bar diagrams. They represent more than one type of data at a time and so two or more bars (as the case may be) are constructed at a time side by side.

Example 13 (a)

The following table gives the number of motor-cars produced in the three countries during the period 1929 –1935.

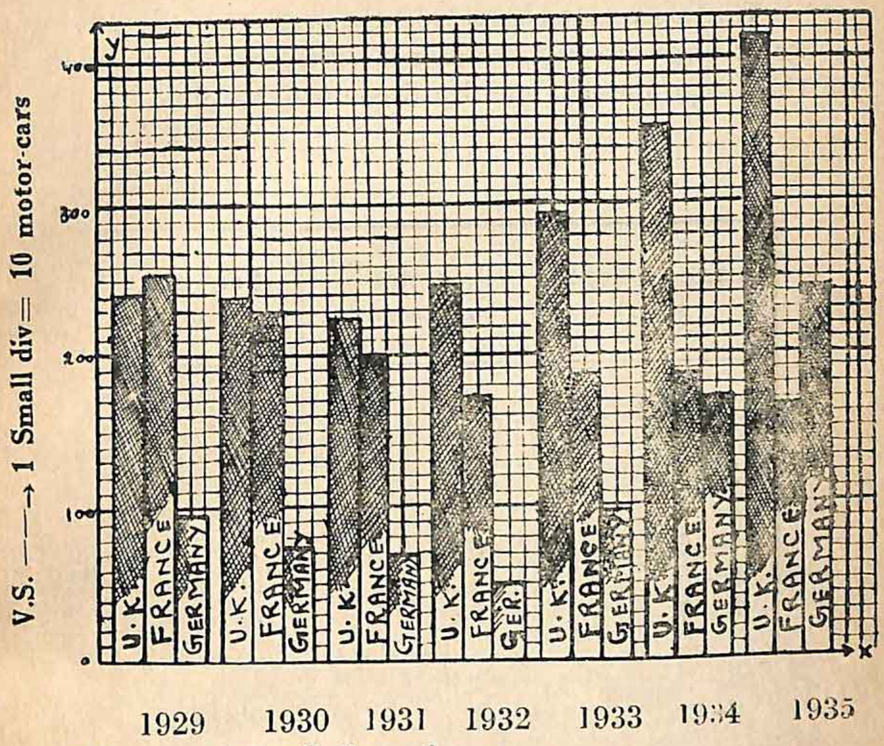
Year	Germany	France	U.K.
1929	96	254	241
1930	74	231	241
1931	68	201	226
1932	50	172	248
1933	99	189	296
1934	172	187	355
1935	245	166	417

Solution:—

Taking three bars at a time, side by side, we completed the diagram. Here the time (year) is taken along the X-axis and the number of cars produced in these countries along the Y-axis.

Diagram No. 2 (a)

"Multiple Bar Diagram for No. of Cars
Produced in U K.,France & Germany"



H.S. ——→ 6 Small divs. = 1 year
2 Small divs. = width of each bar
1 Small div. = spacing between each set of three bars

Example No. 13 (b):—

The following table provides the percentage of cultivators and percentage of cultivated area in different sizes of holdings in U.P. Depit the data in a bar diagram to the scale ?

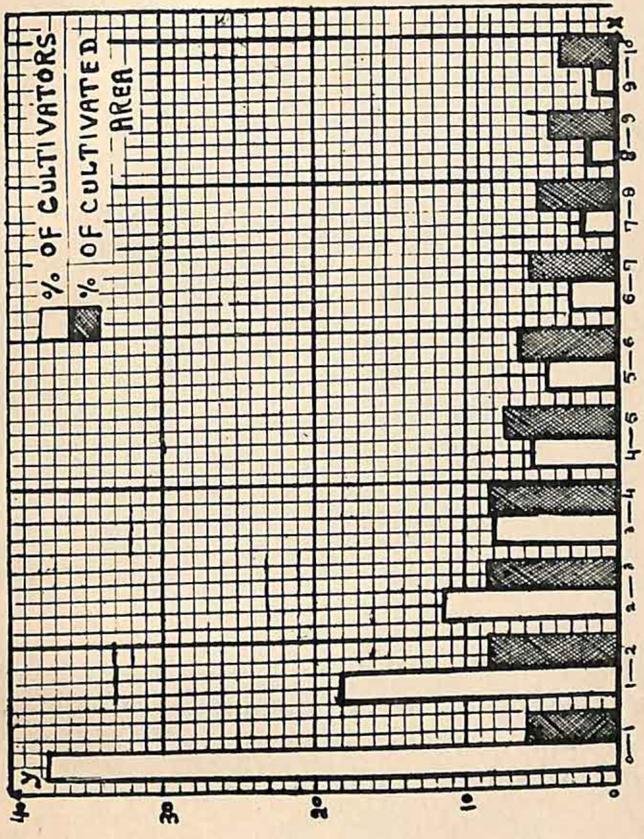
(M.Sc. Ag. 1964)

Size of holding (acres)	Percentage of cultivators	Percentage of area
Up to 1 ..	37·8	6·0
1—2	18·0	8.1
2—3.....	11·6	8·7
3—4.....	8·1	8·4
4—5	5·7	7·6
5—6.....	4·2	6·8
6—7.....	3·0	5·9
7—8.....	2·3	5·1
8—9	1·8	4·4
9–10.....	1·4	3·9

Solution:—

The percentage of cultivators and the cultivated area in different sizes of holdings will be depicted by the compound bar-diagram. Corresponding to each holding size, two abjacent bars will be constructed where one of the two will represent the % of cultivators and the other % of cultivated area.

Diagram No: 2 (b)



H.S. ——→ 4 Small divs = 1 acre (size of holding)
2 " " = Width of each bar
1 " " = Spacing between each pair of bars

V.S. ——→ 1" = 10% of cultivators and area

(3) **Sub-divided Bar Diagrams**:—These diagrams are suitable for the data given in the complex tables; where the magnitudes of the items are sub-divided into sub-classes.

Example (14)

The following table gives the populations of males out of the total populations for different years for a district. Draw a suitable diagram to show the populations given in the data.

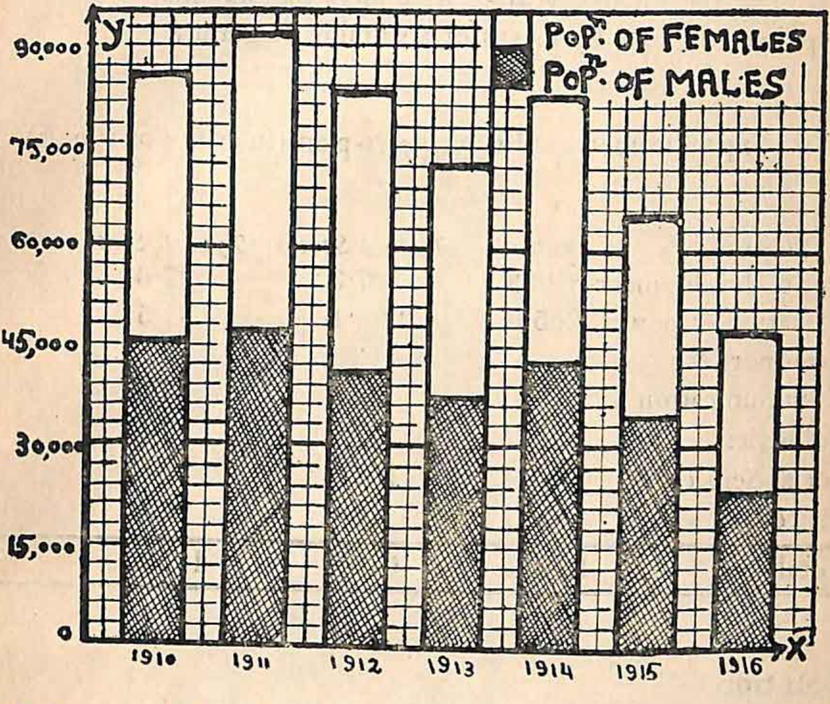
Year	Total Population	Population of Males
1910	85761	45761
1911	92821	47212
1912	83728	41312
1913	72511	37256
1914	83123	42218
1915	65735	34725
1916	46849	23428

Solution:—

The bars are constructed for different years and their lengths are proportional to the magnitudes of the total populations. Then each bar is sub-divided to show the populations of males for the different years.

Diagram No. 3

"Sub-divided bar diagram for total population
& population of males"
from (1910–16)



V.S.——→1 Small div.=3000 population

H.S.——→3 Small divs.=1 year
=width of each bar
2 Small divs.=Spacing between two consecutive bars

Example (15):—

The following table shows the expenditure incurred by the central government and the governments of States of type A & B in the Ist. five years-plan under the six major heads. Represent the data by means of a suitable diagram ?

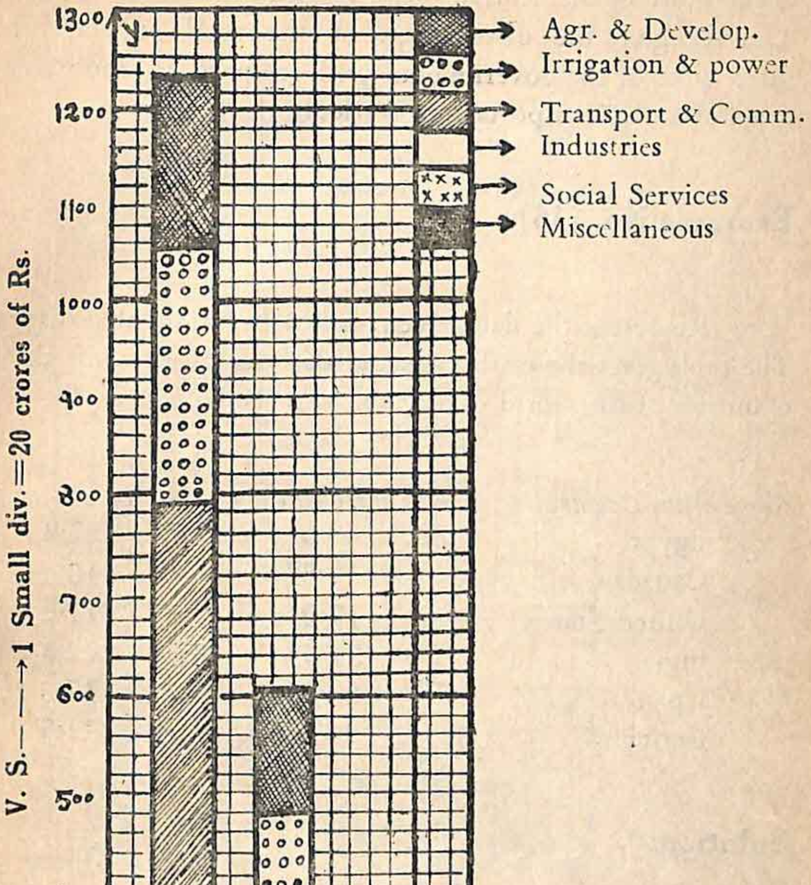
Expenditure in the Ist five years-plan (in crores of rupees)

Subject	*Central*	*Type A States*	*Type B States*
Agr.& development	186·3	127·3	37·6
Irrigation & power	265·9	206·1	81·5
Transport & Communication	409·5	56·5	17·4
Industries	146·7	17·9	7·1
Social Services	191·4	192·3	28·9
Miscellaneous	40·7	10·0	7·0
Totals	1240·5	610·1	173·2

Solution:—

The date can be represented by means of a sub-divided bar diagram. Three bars will be constructed and their heights would be in proportion of the total expenditure of the three types of governments respectively. Further each bar will be sub-divided into six parts and the height of each part will be proportional to the expenditure incurred on it.

Diagram No. 4


V. S.——→1 Small div.=20 crores of Rs.
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Y
X
Agr. & Develop.
Irrigation & power
Transport & Comm.
Industries
Social Services
Miscellaneous
C. GOV. G. of T. A. G. of T. B.

The diagram provides the information about the total expenditure of the central, type A and type B governments. This also gives the distribution of the total expenditure of three types of the governments with respect to the items of expenditure (Transportation, Agriculture etc).

Example No. (16)

Represent the data given below by a suitable diagram. The table gives the birth rates and the death rates of the six countries of the world during the year 1937.

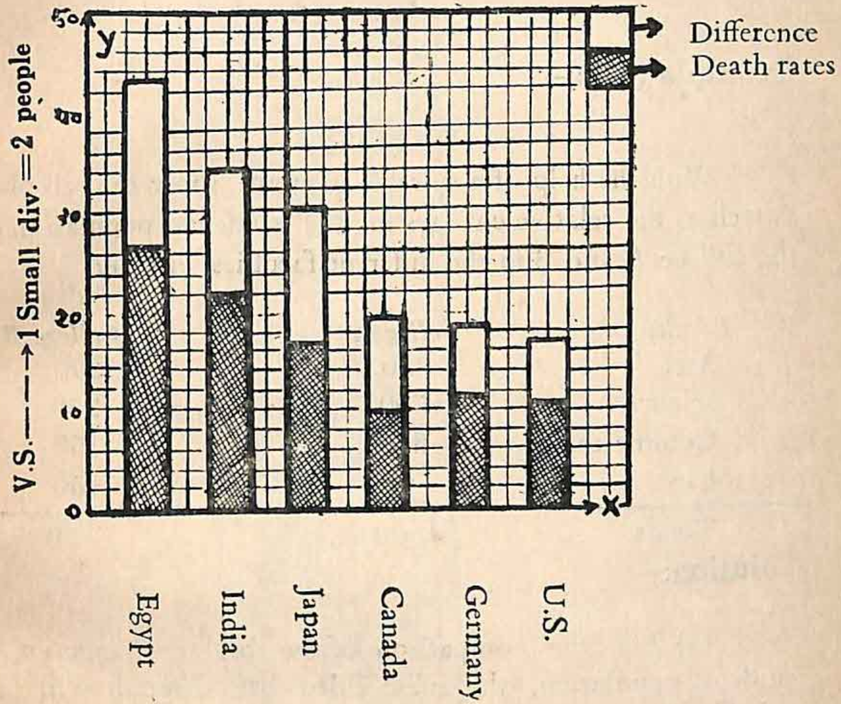
Name of the Country	*Birth rate*	*Death rate*
Egypt	43·5	27·2
Canada	19·8	10·2
United States	17·0	11·2
India	34·5	22·4
Japan	30·6	17·0
Germany	18·8	11·7

Solution:—

These birth and death rates can be represented by sub-divided bar diagrams. Six bars are constructed with their heights proportional to the birth rates. Further the bars are Sub-divided into 2 parts, the lower one shows the death rates and the remaining upper portion shows the difference between the birth and the death rates.

Diagram No. 5

Showing the birth & death rates of six countries in 1937



H.S. ——→2 Small divs. = width of each bar
" =Spacing between two consecutive bars

Percentage sub-divided bar-diagram:—

In this diagram, the total values are taken equal to 100 and the component parts are expressed in percentages. As the length of each bar in this diagram is the same, so it cannot give the comparison between the absolute magnitudes of the components. But the relative changes in the component parts are satisfactorily compared.

Example (17):—

With the help of a suitable diagram, show the absolute as well as the relative changes in the students population of the college A and B in the different faculties in 1947.

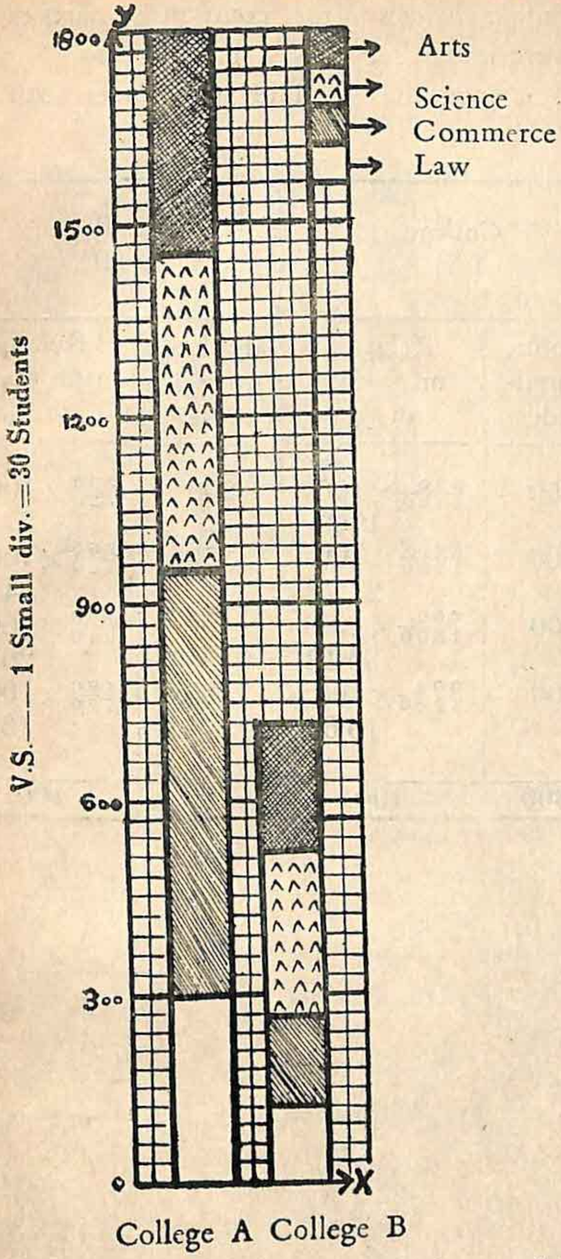
Faculty	*College A*	*College B*
Arts	350	290
Science	500	250
Commerce	650	150
Law	300	150
Totals	1800	720

Solution:—

(a) For the comparison of the absolute changes in the students population, the sub-divided bar diagram will be a suitable diagram.

Diagram No. 6 (a)

h owing the absolute changes in Students population in 1947



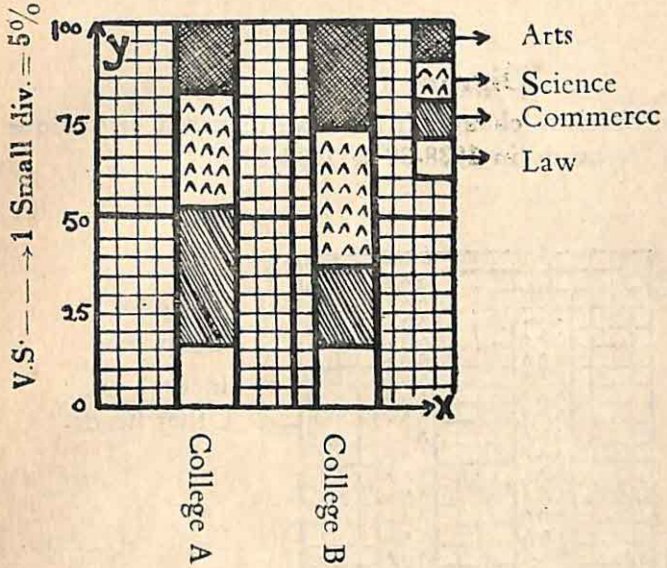
H.S.——→5 Small divs. = width or each bar

(b) The relative changes in the component parts can be shown by the percentage sub-divided bar diagram. For this purpose, first we change the absolute magnitudes into the percentages.

Faculty	College (A)		College (B)	
	absolute magnitude	Relative magnitude in %	absolute magnitude	Relative magnitude in %
Arts	350	$\frac{350}{1800} \times 100 = 19{\cdot}44$	200	$\frac{200}{720} \times 100 = 27{\cdot}77$
Science	500	$\frac{500}{1800} \times 100 = 27{\cdot}77$	250	$\frac{250}{720} \times 100 = 34{\cdot}72$
Commerce	650	$\frac{650}{1800} \times 100 = 36{\cdot}13$	150	$\frac{150}{720} \times 100 = 20{\cdot}85$
Law	300	$\frac{300}{1700} \times 100 = 16{\cdot}66$	120	$\frac{120}{720} \times 100 = 16{\cdot}66$
Totals	1800	100	720	100

Diagram No. 6 (b)

Showing the Relative Change in Students population in 1947



H. S. — — → 3 Small divs. = width of each bar
4 " " = spacing between two bars

Example (18)

For the following data, show by a suitable diagram the comparison between the relative and the absolute changes.

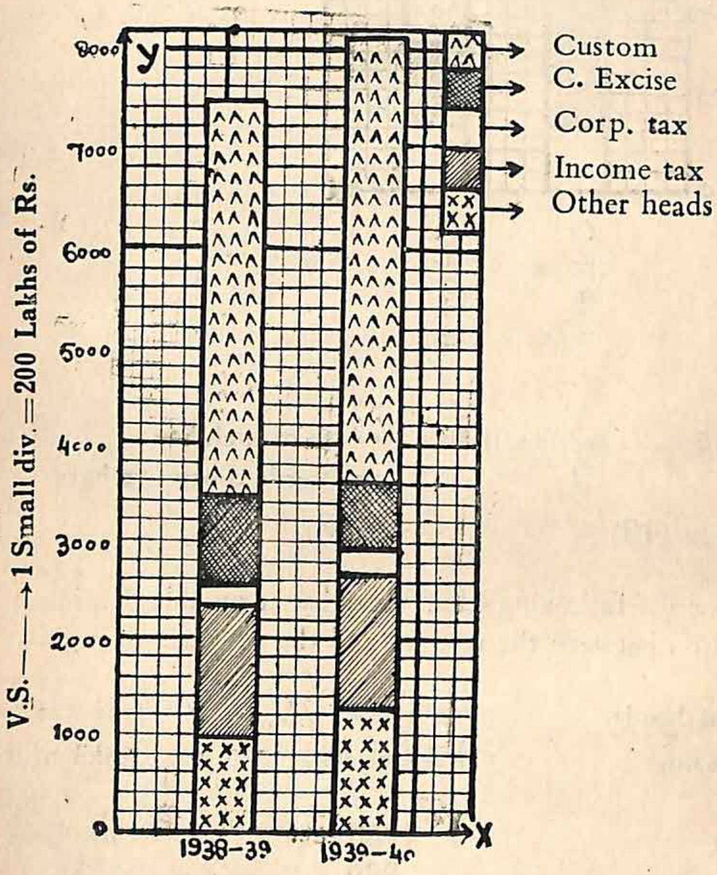
Principal heads of Revenue	Year 1938—39 (Lakh of Rs.)	Year 1939—40 (Lakh of Rs.)
Custom	4050	4588
Central Excise	868	652
Corporation tax	204	238
Incometax	1374	1420
Other heads	974	1262

Solution –

(a) For the comparison of the absolute changes, the subdivided bar diagram will be a suitable diagram.

Digram No. 7(a)

Showing absolute changes in amounts incurred on various heads in 1938-39 & 1939-40



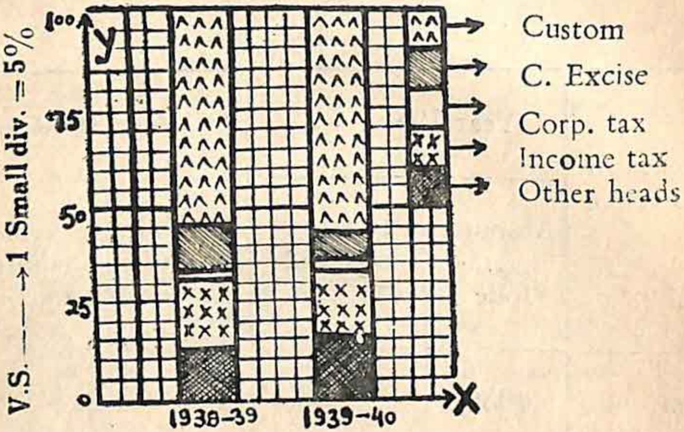
H. S. ——→3 Small divs. = width of each bar = 1 year
4 " " = spacing between two bars

(b) For the relative changes in the component parts, the percentage sub-divided bar diagram will be constructed. For this purpose, we change the absolute magnitudes into the percentages as shown below—

Heads of Revenue	Year 1938—39		Year 1939—40	
	Absolute magnitude	Relative magnitude in %	Absolute magnitude	Relative magnitude in %
Custom	4050	54·3	4588	56·4
Central Excise	868	11·6	652	8·0
Corporation tax	204	2·7	238	2·9
Income tax	1374	18·4	1420	17·5
Other heads	974	13·0	1262	15·2
Totals	7470	100	8160	100

Diagram No 7 (b)

Showing Relative changes in amounts incurred on various heads in 1938—39 & 1939—40



H. S. — — →3 Small divs. = width of each bar = 1 year
4 " " = Spacing between two bars

Bilateral Bars: In this case, the bars are drawn above or below the base line (in the case of vertical bars) and to the left and right perpendicular to the base line (in the case of horizontal bars). The bars above the base line or to the right of the base line are used for the +ve quantity and the bars below the base line or to the left of the base line are used for —ve quantity in the data at hand.

Example (19):—

The following table gives the number of houses in ten small towns of India during the census of 1941 and 1951.

Town :	1	2	3	4	5	6	7	8	9	10
No. of houses in 1941 :	200	300	400	480	520	300	400	280	750	570
No. of houses in 1951 :	250	340	420	495	530	290	380	250	710	520

Represent the increase or decrease in the number of houses in 1951 in comparison to the census of 1941.

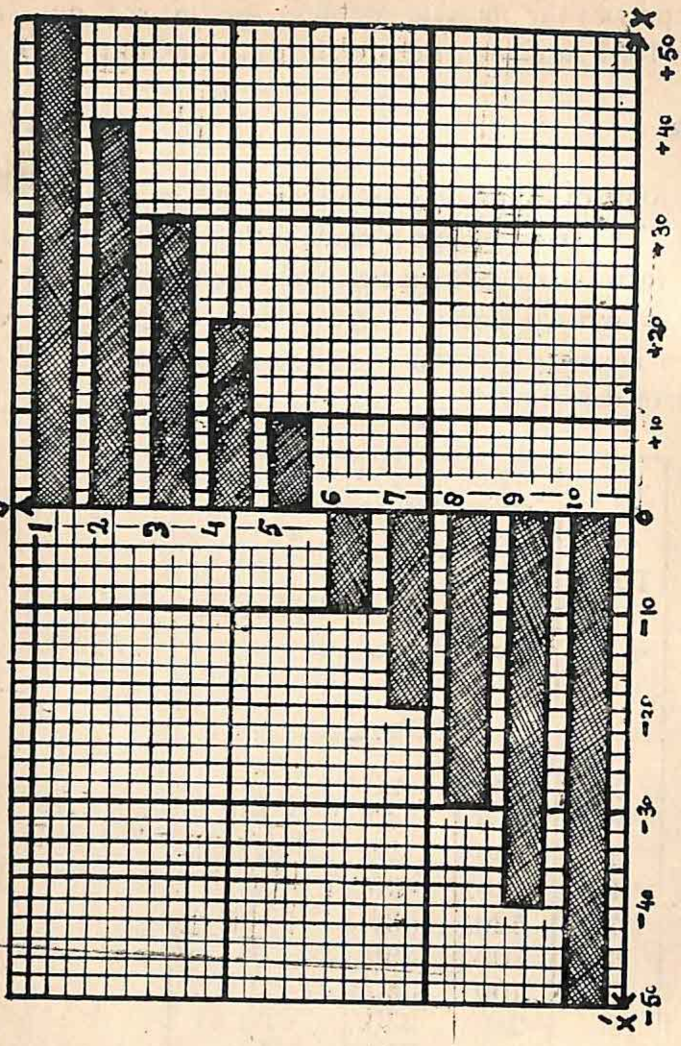
Solution: --

In some cases, there is an increase while in other there is a decrease. The arrangement is made such that the towns which show an increase are written first and the towns which show decrease are written later on. Within these graphs, the towns have been arranged according to the magnitudes of the changes (increase or decrease).

Town	No. of houses in		Difference
	1941	1951	
1	200	250	+50
2	300	340	+40
3	400	420	+20
4	480	495	+15
5	520	530	+10
6	300	290	−10
7	400	380	−20
8	280	250	−30
9	750	710	−40
10	570	520	−50

In the diagram, the +ve changes (increase in the no. of houses) are shown by the bars drawn to the right of the base line and those showing the —ve changes (decrease in the no. of houses) are drawn to the left of the base line.

Diagram No.8



H.S. ——→ 1 Small div.=2 houses

Example No (20):—

Represent the following data by a suitable diagram showing the differences between the proceeds & costs.

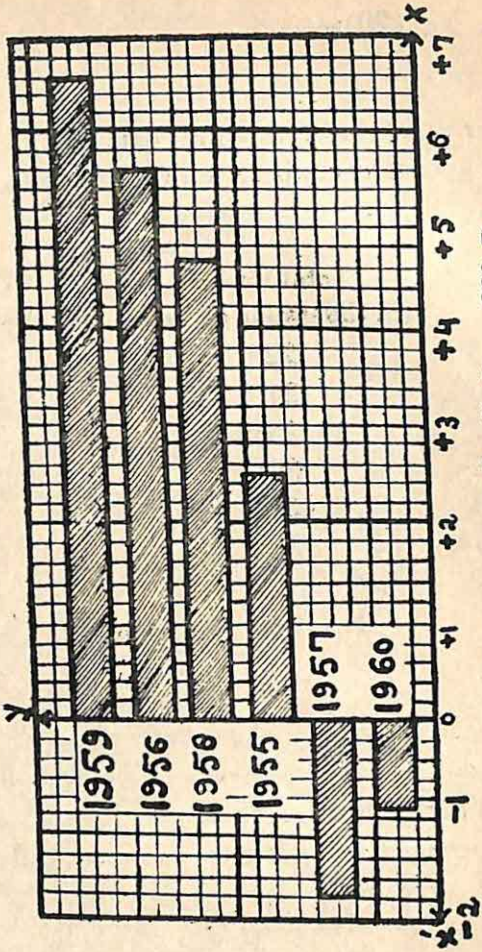
Year	Total Proceeds (in thousands of Rs.)	Total costs (in thousands of Rs.)
1955	22·0	19·5
1956	27·3	21·7
1957	28·2	30·0
1958	30·3	25·6
1959	32·7	26.1
1960	33·3	34·2

Solution:—

First of all we take the differences of the total proceeds & the total costs and then +ve differences are written first and —ve differences after them. Within these groups, they have been arranged according to the magnitudes of the changes (increase or decrease).

Year :	1959	1956	1958	1955	1957	1960
Difference:	+6·6	+5·6	+4·7	+2·5	−·90	−1·8

Diagram No. 9
Differences between Proceeds & Costs
from 1955–60



H. S. ——→1 Small div. = 200 Rs.

(2) **Two Dimensional Diagrams** : (Area Diagrams)

In these diagrams, the areas are proportional to the magnitudes of the data, The common diagram of this type are rectangles, pie and square diagrams. The square and pie diagrams serve the same purpose but the pie diagrams are the easiest to draw and they can be made accurately. Hence the pie diagrams are commonly used in place of square diagrams

(a) Pie/circle diagrams : When the differences between any two quantities to be compared are large, bar diagrams can not be used, as one of them will be extremely small.

and the other large. In such cases, the squares or Pie diagrams are used. For drawing the circles, the squar roots of the quantities are taken as the radii of the circles and so the areas of the circles are proportional to the magnitudes of the data.

(b) Sub-divided Pie diagrams : When we want to compare the totals as well as their components with one another, we may use sub-divided Pie diagrams. The total value is equaled to $2\pi=360°$ and then component parts are expressed in terms of angles. Having determined the angles corresponding to different components, the circle is divided into different sectors on the various angles pre-determind.

Example No. (21):—

Represent the following data by a suitable diagram—

Countries	**Production of cane sugar in 1938-39 in Quintals 0000's omitted**
1. India	2750
2. Java	1550
3. Hawai	835
4. Columbia	51

Solution:—

Here the difference between the quantities 51 & 2750 is very large. Thus Pie diagram will be a suitable diagram to represent the data. Now we construct the following table—

Countries	Quintals (1) 0000' S omitted	Square roots (2)	[3] Radii in inches
1. India	2750	52·44	1·05
2. Java	1550	39·37	0 79
3. Hawai	835	28·90	0 58
4. Columbia	51	7·14	0·14

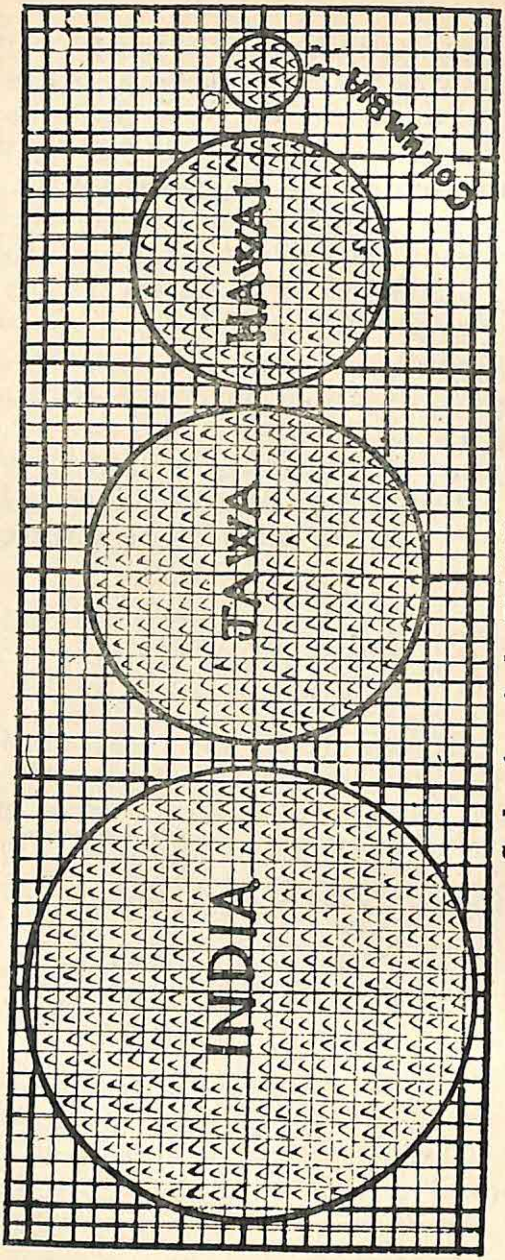
The column (2) gives the square roots of the figures written in column (1) and the column (3) contains the the numbers which are obtained by dividing the numbers of the column (2) by 50 which are the radii in inches. The circles with these radii are arranged in the descending oder. For finding the scale, we calculate the area of the Ist circle= $\frac{22}{7}\times(1\cdot05)^2=3\cdot14\times1\cdot1025$ sq. inche .

Thus $3\cdot14\times1\cdot025$ sq. inches$=$2750,0000 quintals

$\therefore$ 1 sq. inch $=$ 796,5000 quintals.

Diagram No. 10

Production of Cane Sugar in Certain Countries during 1938—39



Scale : 1 sq. inch=796,5000 quintals.

Example 22:—

The allocations for Madras & U.P. States under the second 5 years plan are as given below—

Heads	Madras State (lakhs of Rs.)	U.P. State (lakhs of Rs.)
1. Agr. & Cumunity Development	3535·5	6763·9
2. Irrigation & Power	7125·0	8042·5
3. Industry & mining	1520·0	1643.4
4. Transport	807·5	1723·2
5. Social Services	4065·9	6863·8
6. Miscellaneous	252·1	272·8

Draw a sub-divided Pie diagram to compare the cost of development under each head in the two states?

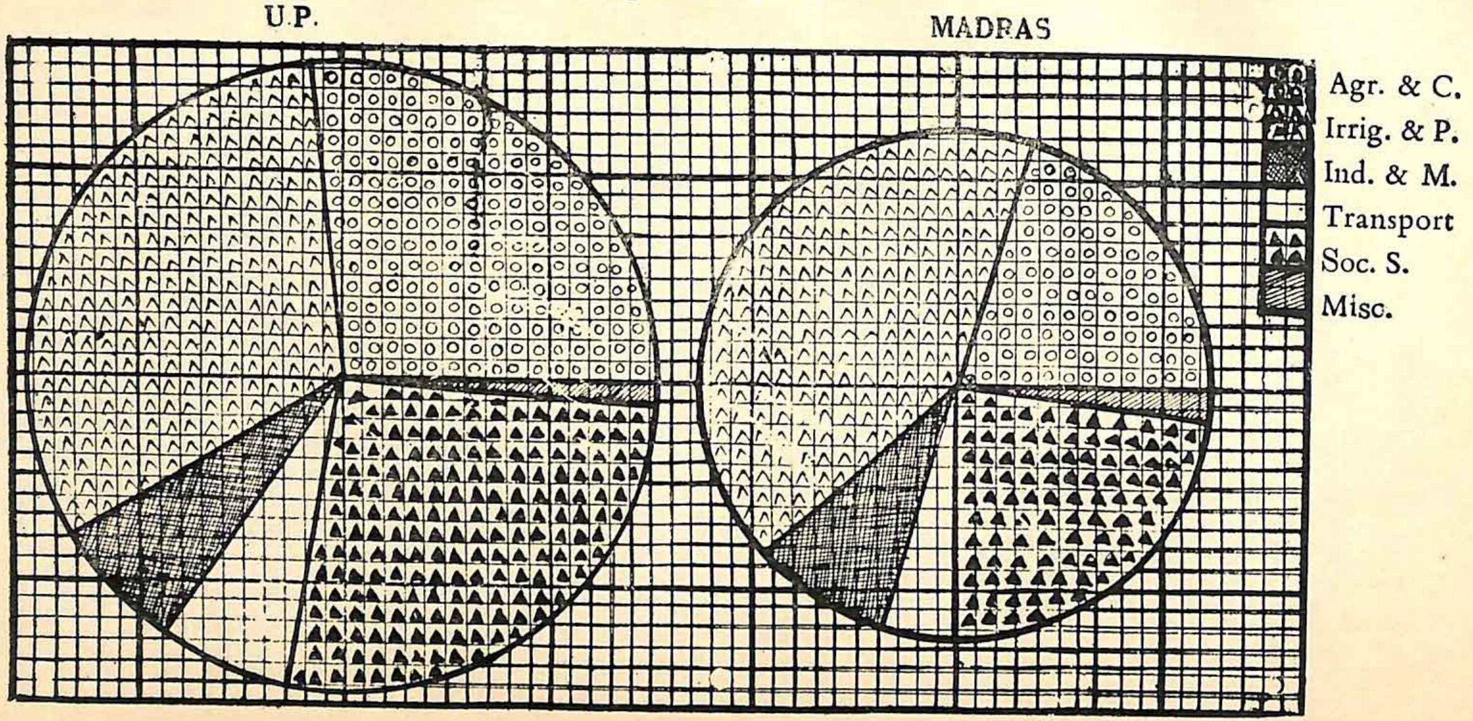
Solution:—

The solution for the problem is given in the tabular form as shown on the next page.

Heads	Madras State		U. P. State		Square roots	Radii in inches (dividing by 100)
	Cost	angle	Cost	angle		
1. Agr. & community development	3535·5	73·5	6763·9	96·04	$\sqrt{17306} = 131·5$	1·3
2. Irrigation & Power	71 5·0	148·2	8042·5	114·20	$\sqrt{25309·6} = 159·09$	1·6
3. Industry & mining	1520·0	31·6	1643·4	23·34		
4. Transport	807·5	16·8	1723·2	24·47		
5. Social services	4065·9	84·6	6863 8	97·47		
6. Miscellaneous	252·1	5·3	272·8	4·48		
Totals	17306·0	360°	25309·6	360°		

Steps to construct the diagram : We take the square roots of 17306·0and 25309·6 which are 131·5 and 159·09 andt han divide these square roots by any number(say 100 here)to obtain the radii of the circles.Then the sectors on various angles are cut from these circles according to the angles shown in the table. In this way,the areas represented by different sectors on a circle are proportional to the magnitudes of the components contained for a state as the main heads. Since $3·14\times(1·3)^2=$ 5·31 equivalents to17306 lakhs of Rs, so Scale is 1 sq. inche=3259·1 Lakhs of Rs.

Diagram No. 11



Example (23):—

Represent the data of the following table by means o a suitable sub-divided Pie diagram—

Clearing house Statistics in 1940–41 and 1947–48 in certain cities—

City	Total amount (Rs.)	
	1940—41	1947—48
Bombay	80,232	255,264
Calcutta	100,853	259,996
Delhi	2,853	12,646
Kanpur	1,920	10,983
Karanchi	4,676	27,481
Lahore	1,633	4,954
Madras	10,865	34,794
Others	4,228	51,896
Totals	2,07,260	658,014

Solution:—

The steps in the construction of Pie diagram are given below—

(i) First of all, calculate the square-roots of the total amount of clearing house returns for the two years i.e. calculate $\sqrt{207260}$ and $\sqrt{658014}$ which are 455 & 811 respectively.

(ii) Divide these quantities (455, 811) by a suitable number to obtain the radii of the circles.

(iii) Then calculate the angles corresponding to the amounts for each city in the separate sessions. The calculations are done in the following tabular manner—

Cities	Year 1940—41 Angles calculated (degree)	Year 1947—48 Angles calculated (degree)
Bombay	$\frac{80232\times 360}{207260}=139{\cdot}4$	$\frac{255264\times 360}{658014}=139{\cdot}7$
Calcutta	$\frac{100853\times 360}{207260}=175{\cdot}2$	$\frac{259996\times 360}{658014}=142{\cdot}3$
Delhi	$\frac{2853\times 360}{207260}=5{\cdot}0$	$\frac{12646\times 360}{658014}=6{\cdot}9$
Kanpur	$\frac{1920\times 360}{07260}=3{\cdot}3$	$\frac{10983\times 360}{658011}=6{\cdot}0$
Karanchi	$\frac{4676\times 360}{207260}=8{\cdot}1$	$\frac{27481\times 360}{658014}=15{\cdot}0$
Lahore	$\frac{1633\times 360}{207260}=2{\cdot}8$	$\frac{4954\times 360}{658014}=2{\cdot}7$
Madras	$\frac{10865\times 360}{207260}=18{\cdot}9$	$\frac{34794\times 360}{658014}=19{\cdot}0$
Others	$\frac{4228\times 360}{207260}=7{\cdot}3$	$\frac{51896\times 360}{658014}=28{\cdot}4$
Totals	360	360

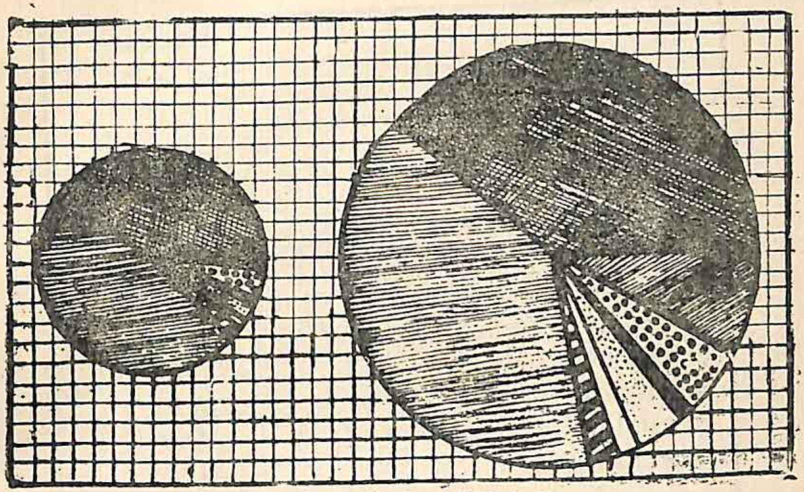
Finally we divide 455 & 811 by a number say 400 to get 1.14 and 2·05 as the approximate radii in inches and the sectors on various angles are cut from these circles. The area enclosed within each sector represents the magnitude of the amount returns in the specified years for different cities.

Scale : 1 sq. inch = 52389 Rs.

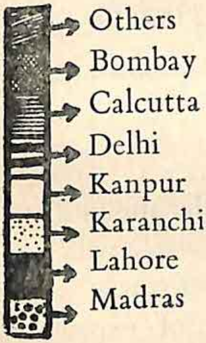
Diagram No. 12

'representing clearing house statistics'

Year : 1940—41 **Year : 1947—48**



Scale : 1 sq. inch = 52389 Rs.



Example 24—

The following table gives the total outlay on rural development proposed in the first Five Years Plan and its break down into major items. Give a suitable diagramatic representation of the data. (M.Sc Ag.Agra, 1965)

Item	Amount (in crores of rupees
Agr. & community development	360·43
Irrigation …………………………	167·97
Irrigation & Power (multipurpose projects)………………	265·90
Power………………………………	127·54
Transport & communications……	497·10
Industry …………………………	173·04
Social services…………………	339·81
Rehabilitation………………	85·00
Miscellaneous ……………………	51·99
Total	2,068·78

Solution:—

The suitable diagram for representing the above data is sub-divided Pie diagram. The total area of the circle will represent the total amount i.e. 2068·78 crores of rupees and its components will represent the various items of expenditure under the plan.

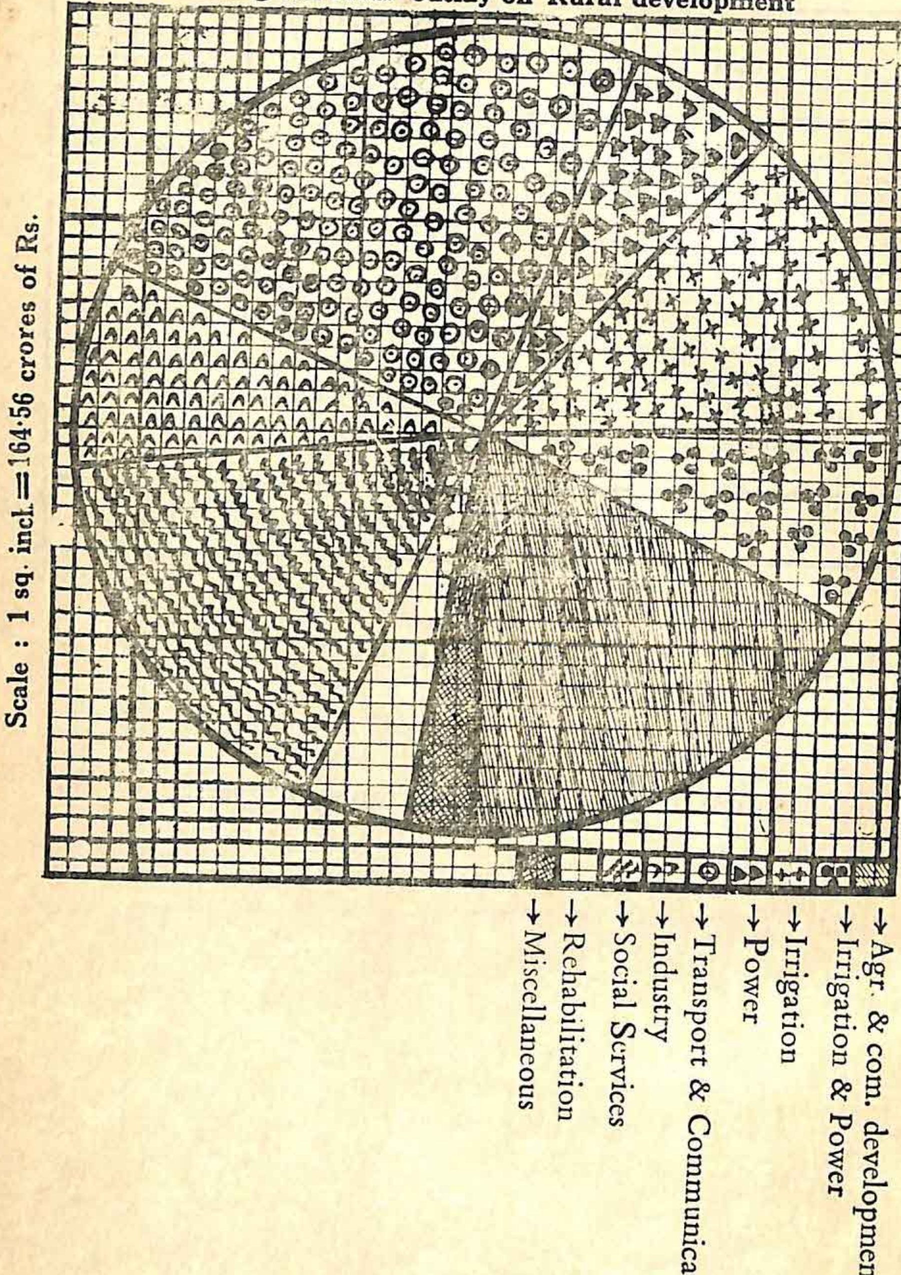
Item	Amount (in crores of Rupees)	Angle
Agr. & com. development	360·43	62·7°
Irrigation & power	167·97	29·2°
Irrigation	265·90	46·3°
Power	127·54	22·2°
Transport & communication	497·10	86·5°
Industry	137·04	30·1°
Social services	339·81	59·1°
Rehabilitation	85·00	14·8°
Miscellaneous	51·99	9·0°
Totals	2068·78	359·9° ≈ 360

Let the radius of the circle be 2″, then πr^2 = Area or $\frac{22}{7} \times 4$ sq. inches = 2068·78 crores of rupees.

$\therefore$ Scale : 1 sq. inch = 164·56 crores of rupees.

Diagram No. 13

Showing the total outlay on Rural development



Rectangles : The rectangles are the two dimensional diagrams and are used when the two magnitudes which are related to a third one are to be represented on the same diagram. For example, the average yield of wheat multiplied by the total acreage under wheat represents the total production of wheat. To represent it diagramatically, one side of the rectangle is taken proportional to the average-yield and the other is proportional to the total acreage; the area of the rectangle gives the totat production of wheat. Sub-divided rectangles are used when one of the magnitudes is divided into several components.

Example No (25)

An analysis of the monthly wages paid to workers in the two firms A & B gives the following results—

	Firm A	Firm B
No. of wage earners	80	100
Average monthly wage (Rs.) :	52.5	47·5

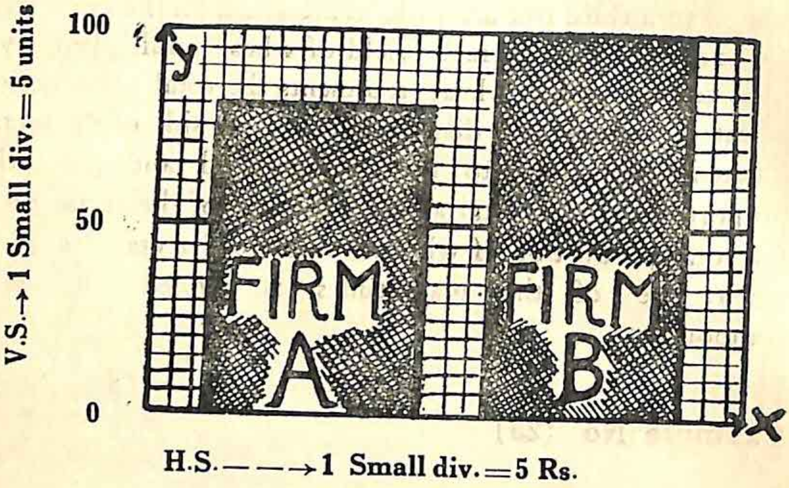
Represent the above facts diagramatically ?

Solution:—

Here the average wage multiplied by the total number of workers gives the daily paid roll. Hence the most suitable diagram for representing the above data is to construct the two rectangles, one for each firm. One side of the rectangles will be proportional to the average wage and the other will be proportional to the total no. of workers employed in a firm.

Diagram No. 14

Showing the average monthly wage and number of wage earners



The family budget data, where we compare the absolute totals as well as the percentages of the various items to the total, is very suitably represented by the rectangles. In this case, the height of each rectangle is kept 100 and the width is made proportional to the total magnitude.

Example No. 26

The following table gives the details of the monthly expenditure of three families. Represent them by a suitable diagram—

Item of the expenditure	**Family A** Rs. An.	**Family B** Rs. An.	**Family C** Rs. An.
Food	12—0	25—0	28—0
Clothing	2—8	8—0	14—0
House rent	2—0	4—0	7—
Education	1—0	5—0	7—0
Miscellaneous	2—8	8—0	14—0
Totals	20—0	50—0	70—0

Solution:—

First we prepare the following table:—

Item of expenditure	Family A		Family B		Family C	
	Actual expenses	%	Actual expenses	%	Actual expenses	%
Food	12·00	60·00	25·00	50·00	28	40.00
Clothing	2·50	12·50	8·00	16·00	14	20.00
House rest	2·00	10·00	4·00	8·00	7	10.00
Education	1·00	5·00	5·00	10·00	7	10.00
Miscellaneous	2·50	12·50	8·00	16·00	14	20.00
Total	20·00	100	50	100	70	100

The heights of the three rectangles will be taken as 100 for each and their width in the ratio 20:50:70. Each rectangle will be further divided into component parts according to the figures in the percentage column.

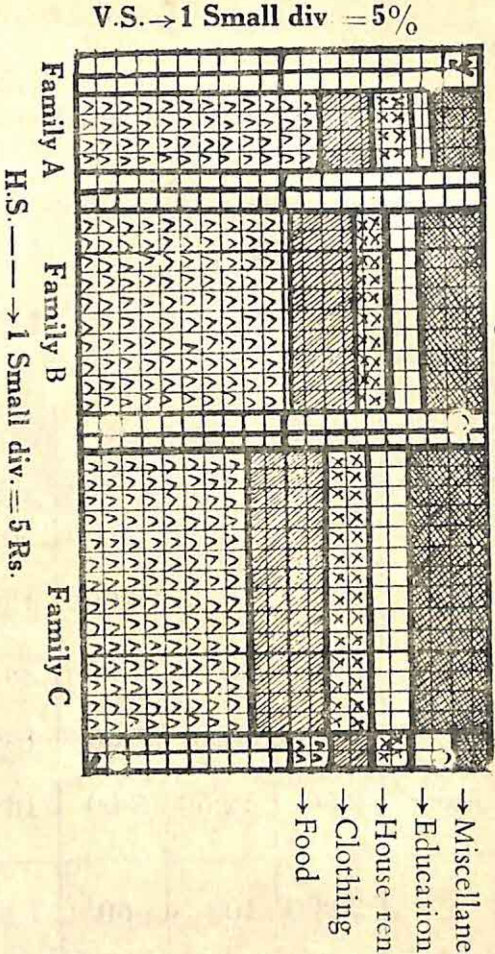


Diagram No: 15
Showing the family-budget data

The above diagram represents—

(i) The actual expenses on each item by the area of the corresponding component part of the rectangle.

(2) The % of the item to the total by the height of the corresponding component part of the rectangle.

(3) The total expenses for the family by the whole area of the rectangle.

Example (16):—

Represent the following data by a suitable diagram—

Year	Average yield Md/acre	Total yield	Area under crop		Total (acreage)
			Irrigated	Non-Irrigated	
1940–41	9.5	427.5	22	23	45
1941–42	8.4	361.2	23	20	43

Solution:—

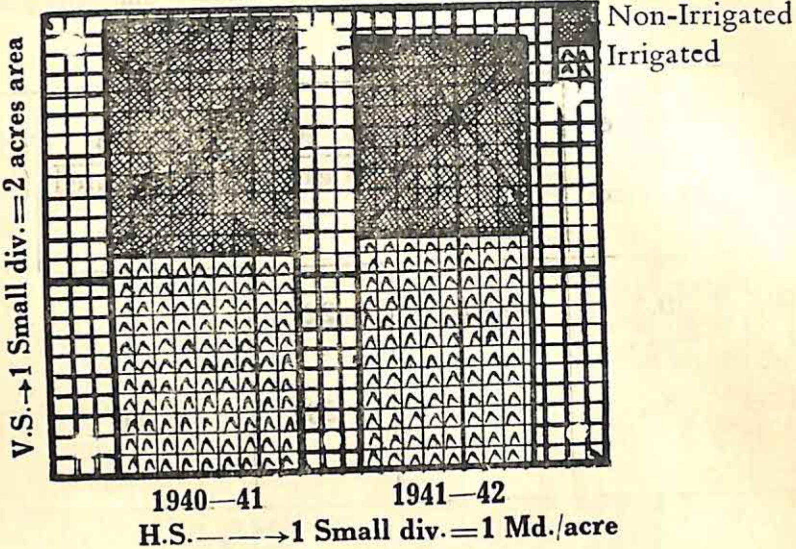
Here we note the relation—

Average yield × total area = total yield.

So the rectangles are the suitable diagrams to represent the given data. Here we construct the two rectangles on the same horizontal axis taking their width in the ratio 9·5:8·4 and heights in the ratio 45:43. Each rectangle will be further divided into two components according to the irrigated and non-irrigated areas.

Diagram No. 16

Showing the average yield md/acre & area under crop irrigated and non-irrigated



Three dimensional Diagrams : When the ratio between the two quantities to be compared is so large that the two diamentional diagrams are not suitable for representing them, we may represent them by three dimensional diagrams. The volumes of the diagrams represent the magnitudes of the quantities to be compared. In this type of diagrams the spheres, cubes, and the prisms are common but cubes are generally drawn in practice.

Chapter III

Measures of Central tendency (averages) and dispersion

The histogram or frequency curve gives the general idea of the distribution of the variate and hence the frequency graph can be used to study and compare the given distributions. But the study through the graphs depends upon the accuracy and skill of the eye, which is an uncertain factor. Thus the study and comparisons by it may not be very reliable. Therefore it is necessary to know certain features of the distribution, which give an idea of the distribution and can be determined arithmatically. Two such features of the distribution are its average and dispersion (variation).

Average : An average is the value of the variate which claims to represent the distribution. Some of the variate values will be above this value and others below it and so it is known as a measure of the central tendency.

There are three averages (i) Mean (ii) Mode and (iii) Median.

Exp. No. 1

What are the properties of an ideal average. In what circumstances, would you consider the mean, mode and the median, the most suitable statistics to describe the central tendency of the distribution ?

Sol:—

An ideal average should have the following properties:—

(1) It should be rigidly defined and its value should be

definite which can be determined mathematically

(2) It should be based on all observations

(3) Its calculation should not be lengthly and tedious.

(4) It should be least affected by sampling fluctuations.

(5) It must be capable of algebraic treatments i.e. if the averages of the component series are known then the average of the whole series should be expressible in terms of the averages of the component series.

Relative merits & Demerits of different averages :

(1) Mean (Arithmatic average)

Advantages:—

(1) It is readily understood and well defined.

(2) It is based on all observations.

(3) It is easy to calculate.

(4) In most of the cases, it is not affected by the sampling fluctuations.

(5) It is capable of algeberaic treatments.

(6) It gives weights to all items which are directly proportional to their sizes.

Disadvantages:—

(1) It some times gives values which may not be physically possible e.g. the average number of eggs laid by a hen as 18·5 per month.

(2) It gives undue weights to the extreme items.

(3) It can not be calculated in the cases where the extreme ends are open.

(2) **Mode** :

Advantages:—

(1) In most of the cases, it is easily found.

(2) It can be found out from the graph merely by an eye inspection.

(3) It can be found out for the distributions where the ends are open.

(4) It is the type that to the ordinary mind, seems to be the best to represent the group.

Disadvantages:—

(1) No weights are given to the extreme items.

(2) A clearly defined mode does not always exist

(3) It is not capable of algebraic-treatments.

(3) Median :

Advantages:—

(1) It is well defined and can be calculated easily.

(2) It does not give undue weights to the extreme items.

(3) It is possible to calculate even in the cases where the intervals are open.

Exp. No. 2

What is an arithmatic average ? What are its properties ? How would you calculate the Arithmatic Average ?

Sol:—

The arithmatic mean of a series of individuals is obtained by adding up all the values and then dividing the total by the number of individuals.

For example, if the heights of 10 randomly chosen plants are–52″, 55″, 57″, 61″, 64″ 65″, 67″, 68″, 70″, & 71″, then the arithmatic mean or simply the mean of this sample of 10 plants will be the sum $(52+55+\ldots\ldots\ldots+71=630)$ divided by 10. i.e. mean$=\frac{630}{10}=63''$

In general, if $x_1, x_2, \ldots\ldots\ldots, x_n$ are the values of n measurements, the arithmatic mean $\overline{X}$ of the X_s is defined by the relation—

$$\overline{X}=\frac{x_1+x_2+\ldots\ldots+x_n}{n}=\frac{\Sigma x}{n}\ldots\ldots\ldots\ldots\ldots(i)$$

where $x_1, x_2, \ldots\ldots, x_n$ are the given values.

n = number of items

Σ = the sum of

$\overline{X}$ = Arithmatic mean of X_s

Now, the formula (i) will be read as "The arithmatic mean of x'_s is the sum of x'_s divided by their number."

Exp No. 3

The yield of rices in 12 equal plots of a village are given as follows (in maunds)—

68, 80, 94, 104, 120, 114
125, 130, 130, 140, 141, 145, find the mean yield ?

Solution:—

If the yields are denoted by x_s, then

$$\bar{X}=\frac{\Sigma x}{n}=\frac{68+80+\ldots\ldots+145}{12}=\frac{1392}{12}=116 \text{ maunds/plot.}$$

If the value x_1 occurs f_1 times, x_2 occurs f_2 times, and so on, then

$$\bar{X}=\frac{f_1x_1+f_2x_2+\ldots\ldots+f_nx_n}{f_1+f_2+\ldots\ldots+f_n}=\frac{\Sigma fx}{\Sigma f}=\frac{\Sigma fx}{n}\ \ldots\ldots\ldots\text{(ii)}$$

as $\Sigma f = n$

where $\bar{X}$=Arithmatic average and $x_1, x_2, \ldots\ldots x_n$ are the given values of the variate x_1.

$f_1, f_2, \ldots\ldots f_n$ are the respective frequencies of $x_1, x_2, \ldots\ldots x_n$.

$n=f_1+f_2+\ldots\ldots+f_n=\Sigma f$

To make the procedure clear, let us take the yield of milk by 80 cows of a dairy farm on a certain day (in kilograms)—

Milk-yield :	11	14	17	21	25	30	32
No. of cows	6	9	12	16	21	9	7.

To calculate the arithmatic means of the milk yield, we prepare the following table—

Milk-yield (1) X (in Kgm.)	No. of cows (2) f	(3) fx
11	6	66
14	9	126
17	12	204
21	16	336
25	21	525
30	9	270
32	7	224
Totals	$80=\Sigma f$	$1751=\Sigma fx$

In column (1), we put the variate values and in column (2) their respective frequencies against them. In column (3) we put the product of (1) & (2). For example, the first value 66 of the col^n. (3) is the product of 11 & 6 which are the Ist. figures in the col^n. (1) & (2) respectively. Similarly, the other figures of col^n. (3) are attained. Thus, we get—

$$\overline{X}=\frac{\Sigma fx}{\Sigma f}=\frac{\text{sum of col}^n.\ (3)}{\text{sum of col}^n.\ (2)}=\frac{1751}{80}=21{\cdot}8875 \text{ kgms.}$$

This formula can be applied in a grouped frequency distribution after the classes have been replaced by their respective mid-values. The procedure of calculation will be clear from the following Exp.—

Exp. No. 4

Given the following distribution, calculate the mean?

Height of plants (in cms.)	No. of plants	Height of plants (in cms.)	No. of plants
12·5—17·5	2	37·5—42·5	4
17·5—22·5	22	42·5—47·5	6
22·5—27·5	19	47·5—52·5	1
27·5—32·5	14	52·5—57·5	1
32·5—37·5	3		

Solution:—

The mean is given by the formula $X=\frac{\Sigma fx}{n}$, where the computations of $\Sigma fx = n$ are given below in the tabular form—

(1) x	(2) f	(3) fx
15	2	30
20	22	440
25	19	475
30	14	420
35	3	105
40	4	160
45	6	270
50	1	50
55	1	55
Totals	$72=\Sigma f$	$2005=\Sigma fx$

In the col^n. (1) we have put the mid-values of the classes and in (2) the class frequencies respectively. Thus, the product of (1) & (2) are put in the col^n. (3)

Therefore, Σfx is the sum of all figures in the col^n. (3) and $\Sigma f = n$. the sum fo all figures in the col.(2).Hence we have

$$\overline{X} = \frac{\Sigma fx}{n} = \frac{2005}{72} = 27{\cdot}85 \text{ centimetres.}$$

For a short cut method for computing the arithmatic-mean, the formula is—

$$\overline{X} = A + \frac{\Sigma fd}{n}$$

where A is the assumed mean and

$d = (x - A)$ the deviation of the variate from the assumed mean.

Properties of the Arithmatic Mean:

(1) The sum of the deviations of the velues of x_s from their mean $\overline{X}$ is always zero i.e. if $x_1, x_2, \ldots\ldots x_n$ be the variate values and $\overline{X}$ be their mean, then $(x_1 - \overline{X}), (x_2 - \overline{X}), \ldots (x_n - \overline{X})$ etc. are their deviations form the mean and the sum of deviations is zero. i.e. $\Sigma(x - \overline{X}) = 0$

(2) If $\overline{X}_1, \overline{X}_2, \ldots\ldots\ldots \overline{X}_k$ be the arithmatic means of k distributions with respective frequencies $n_1, n_2, \ldots\ldots n_k$, then the mean $\overline{X}$ of the whole distribution is given by—

$$\overline{X} = \frac{n_1\overline{X}_1 + \ldots\ldots + n_k\overline{X}_k}{n_1 + \ldots\ldots n_k} = \frac{\Sigma n\overline{X}}{\Sigma n}$$

(3) If $x_1, x_2 \ldots\ldots x_n$ be the variate values with their mean $\overline{X}$ and a variate $y = ax + b$ is obtained, then $\overline{Y} = a\overline{X} + b$, where a & b are constants.

Exp. No. 5

If $x_1, x_2, \ldots\ldots x_n$ be the 'n' values of any measurement with their respective frequencies $f_1, f_2, \ldots\ldots f_n$ and mean $\overline{X}$, then show that

$$\overline{X}=A+\frac{\Sigma fd}{n},$$

where A is the assumed mean and

d is the deviation of x from A

i.e. $d=(x-A)$

Sol –

We define $\overline{X}=\frac{\Sigma fx}{n}$

$$=\frac{\Sigma f(x-A+A)}{n}$$

$$=\frac{\Sigma f(x-A)}{n}+\frac{A\Sigma f}{n}$$

$$=\frac{\Sigma f(x-A)}{n}+A.\frac{n}{n}$$

$$\therefore\quad \overline{X}=A+\frac{\Sigma f(x-A)}{n}\ldots\ldots\ldots\ldots\ldots\ldots(i)$$

$$\text{or } \overline{X}=A+\frac{\Sigma fd}{n}$$

In future, for the computation of mean, we shall apply the above (i) formula which reduces the bulk of calculations and hence called the *short cut formula for mean.*

Exp. No. 6

Using the short cut method, compute the mean of the following data–

(a)

No. of branches/ plant (x) :	5	10	15	20	25	30	35	40	45	50
No. of plants (f) :	20	43	75	67	72	45	39	9	8	6

(b)

Age (years)	0–10,	10–20,	20–30,	30–50,	50–80
No. of pesrons :	61,	49	40,	60,	23

Solution:—

(a) To calculate the mean of this data, first we perform the following table—

A=25

(1) x	(2) f	(3) d=x—A	(4 fd
5	20	—20	—400
10	43	—15	—645
15	75	—10	—750
20	67	—5	—335
25	72	0	0
30	45	5	225
35	39	10	390
40	9	15	135
45	8	20	160
50	6	25	150
Totals	384=Σf	—	-1070= Σfd

The assumed mean i.e. A should be taken at such a value of x which divides the whole distribution approximately into two equal parts, so that some of the 'd' values be —ve and the others be +ve. Here, we have taken A=25 and col^n. (3) is obtained by subtracting A=25 from the figures of col^n. (1) as the first value in col^n. (3) is 5—25=—20 and so on, the others. The col^n. (4) is obtained as the respective product of (2) & (3). Then we compute

$$\overline{X}=A+\frac{\Sigma fd}{n}$$

$$=25+\frac{-1070}{384}=25-2{\cdot}8=22{\cdot}2$$

$\therefore$ $\overline{X}=22{\cdot}2$ no. of branches/plant.

(b) To calculate the mean of this data, the following table is performed—

$$A=25$$

(1) x	(2) f	(3) d=x–A	(4) fd
5	61	–20	–1220
15	49	–10	–490
25	40	0	0
40	60	15	900
65	23	40	920
totals	233=Σf		110=Σfd

The col^n. (1) of this table has been obtained dy replacing the classes by their respective mid-values and col^n. (3) is the result of subtracting A=25 (assumed mean) from the figures of col^n. (1). The col^n. (4) is the product of (2) & (3) col^{n}s. with their respective figures. Then we compute the mean

$$\bar{X}=A+\frac{\Sigma fd}{n}.$$

$$=25+\frac{110}{233}=25+0{\cdot}47$$

$$\therefore\ \bar{X}=25{\cdot}47 \text{ years.}$$

Example No. 7

What is median ? Illustrate by an example, how would you calculate the median ?

Solution:—

Median : The median is defined as the value of a variate which divides the distribution into two equal parts. Thus, half of the values lie below the median and half above it.

Calculation of Median :

I—(a) Simple series of odd items :

In a simple series, the variate does not repeat itself. In such a series, if the number of items is odd, the computation of the median is very easy.

First we arrange the values either in the ascending or in the descending order of their magnitudes and then find the size of $\left(\frac{n+1}{2}\right)^{th}$ item in this series. It will be our desired value of the median.

For illustration , consider the following example of 9 items with magnitudes 2, 5, 7, 3, 11, 9, 8, 4 & 6.

To find the median, the items are arranged in the ascending order of their magnitudes i. e.

2, 3, 4, 5, 6, 7, 8, 9, 11, then size of $\left(\frac{n+1}{2}\right)^{th}$ item gives the median. Therefore

$$Md = \text{size of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

$$= \text{size of } \left(\frac{9+1}{2}\right)^{th} \text{ item} = \text{size of } 5^{th} \text{ item}$$

$$= 6$$

$\therefore$ Median $= 6$

I—(b) Simple series of even items :

In this case, the median is the simple arithmatic average of the size of $\left(\frac{n}{2}\right)^{th}$ item and $\left(\frac{n}{2}+1\right)^{th}$ item when the values have been arranged either in the ascending or in the descending order of their magnitudes. For example, take the case of finding the median from the data of eight (even number) values—

62, 65, 67, 64, 72, 53, 59, 55

The ascending arrangement is

53, 55, 59, 62, 64, 65, 67, 72. Then the size of

$\left(\frac{n}{2}\right)^{th}$ item= size of $\left(\frac{8}{2}\right)^{th}$ i.e. 4th item=62

& size of $\left(\frac{n}{2}+1\right)^{th}$ item=size of 5th item =64

Hence, Md.=Simple arithmatic average of the size of $\left(\frac{n}{2}\right)^{th}$ & $\left(\frac{n}{2}+1\right)^{th}$ items.

$$=\frac{62+64}{2}=63$$

$\therefore$ Median =63

Ungrouped data of discrete variate:

If a variate x_1 repeats f_1 times, x_2 repeats f_2 times and so on i.e. if we deal with a frequency distribution of the following type—

Variate value (x)	**frequency (f)**
x_1	f_1
x_2	f_2
x_3	f_3
$\vdots$	$\vdots$
x_n	f_n

, then there are following steps in the calculation of the median.

(i) Arrange the items either in ascending or descending order of their magnitudes, if they are not so. But usually in such a case the data is already arranged and a fresh arrangement is rarely required.

(ii) Compute the cumulative frequencies of the variate values.

(iii) locate the median item which is $\left(\frac{n+1}{2}\right)^{th}$ where n is the total frequency.

(iv) Find the value of the median which is the size of the $\left(\frac{n+1}{2}\right)^{th}$ item.

Consider the example given below to compute the median—

x : 1 2 3 4 5 6

f : 20 15 17 19 13 15

For the computation of the median, we proceed as follows—

x	f	c·f·
1	20	20
2	15	35
3	17	52
4	19	71
5	13	84
6	15	99

Since the median is the size of the $\left(\frac{n+1}{2}\right)^{th}$ item, the median here, is the size of $\left(\frac{99+1}{2}\right)^{th} = 50^{th}$ item.

By inspecting the col^n. of cumulative frequencies we note that the 50th item falls opposite the variate value 3 and so the median here is 3.

In case, when the total of all frequencies is an even number, the median will be the mean of the size of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ items.

3—Grouped data of discrete & continuous variate:

In the case of the grouped data, the arrangement of the items according to their magnitudes is already done. What is equired next to be done, is to calculate.

(i) The cumulative frequencies,

(ii) To locate the median-class, which is $\left\{\frac{n+1}{2}\right\}^{th}$ c.f.

(iii) To determine the value of the median by applying the formula—

$$Md=L+\frac{m-c}{f}\times i$$

where Md– –→median, $m=\frac{n+1}{2}$

L——→lower limit of the median class
i– –→class-interval of the median class,
f——→frequency of the median class
c– –→cumulative frequency of the class following the median class

For illustration, consider the following frequency distribution to find out the median—

class	frequency (f)
0—5	4
5—10	6
10—15	10
15—20	16
20—25	12
25—30	8
30—35	4

To calculate the median for the above data, first we prepare the following table for c.f.

class	frequency (f)	c·f·
0—5	4	4
5—10	6	10
10—15	10	20
15—20	16	36
20—25	12	48
25—30	8	56
30—35	4	60

Here, we want the value of $\left(\frac{60+1}{2}\right)^{th}=30.5^{th}$ item, which is located in the class 15—20. Then the median is determined by the formula—

$$Md=L+\frac{m-c}{f}\times i \qquad \text{where } m=\frac{n+1}{2}=30.5$$

$$=15+\frac{30{\cdot}5-20}{16}\times 5 \qquad L=15$$

$$c=20$$

$$=12+3{\cdot}28 \qquad f=16$$

∴ median=18·28 approximately. $\qquad i=15$

Example No. (8)

What do you mean by the Quartiles ? Calculate the median, the upper and the lower quartiles from the data given below—

% of recovery of sugar on cane	No. of factories	% of recovery of sugar on cane	No. of factories
8.0—8.2	2	9.4—9.6	10
8.2—8.4	5	9.6—9.8	7
8.4—8.6	4	9.8—10.0	6
8.6—8.8	11	10.0—10.2	3
8.8—9.0	11	10.2—10.4	1
9.0—9.2	11	10.4—10.6	1
9.2—9.4	13		
Total			85

Solution—

As the median is such a value of the variate which divides the whole distribution in such a way that half of the observations lie below it and the remaining half above it when the data is arranged in the ascending or descending order of magnitudes.

Similarly, the lower quartile denoted by Q_1 is the **value** of the variate which divides the distribution in such a manner that one quarter of the observations lie below it and and the remaining 3 quarters above it when the values have been arranged in the ascending order of their magnitudes.

The upper quartile denoted by Q_3 is the value of the variate which divides the distribution in such a way that 3 quarters of the observations lie below it and 1 quarter **above** it when the data have been arranged in the ascending order of their magnitudes.

Clearly Q_2 is the median Q_1, Q_3 and median, are also called the *partitioning values of the distribution.* Other partitioning values are Pertiles, Deciles and Hectiles.

The mathematical formula for the computation of quartiles for the grouped data are—

$$Q_1 = L + \frac{q_1 - c}{f} \times i$$

where L--→lower limit of Q_1 class

$$q_1 = \frac{n+1}{4}$$

c—→c.f. of the class following the Q_1 class

f—→frequency of the Q_1 class

i—→class interval of the Q_1 class

and

$$Q_3 = L + \frac{q_3 - c}{f} \times i$$

where $q_3 = \frac{3(n+1)}{4}$

L——→ lower limit of Q_3 class
f- →frequency of Q_3 class
i——→class interval of Q_3 class
c——→c.f. of the class following the Q_3 class

The Q_1 class lies opposite of $\left(\frac{n+1}{4}\right)$ c.f. and Q_3 class lies opposite of $\frac{3(n+1)}{4}$ c.f. For the present example, we have the following table—

% recovery of sugur on cane x	No. of factories (f)	c. f .
8·0–8·2	2	2
8·2—8·4	5	7
8·4—8·6	4	11
8·6—8·8	11	22
8·8—9·0	11	33
9·0—9·2	11	44
9·2—9·4	13	57
9·4--9·6	10	67
9·6—9·8	7	74
9·8—10·0	6	80
10·0—10·2	3	83
10·2–10·4	1	84
10·4—10·6	1	85

$$Md = L + \frac{m-c}{f} \times i \qquad \text{where } m = \frac{85+1}{2} = 43$$

Hence the median class is 9·0—9·2 and so $L=9{\cdot}0$, $i=0{\cdot}2$, $f=11$, $c=33$

$$\therefore \quad Md = 9{\cdot}0 + \frac{43-33}{11} \times 0{\cdot}2$$
$$= 9{\cdot}0 + 0{\cdot}1818$$

$\therefore$ median $= 9{\cdot}1818$

The lower quartile $Q_1 = L + \frac{q_1 - c}{f} \times i$

$$\text{where } Q_1 = \frac{85+1}{4} = 21{\cdot}5$$

Hence the lower quartile class is 8·6—8·8 and so $L=8{\cdot}6$ $i=0{\cdot}2$, $f=11$, $c=11$

$$\therefore \quad Q_1 = 8{\cdot}6 + \frac{21{\cdot}5 - 11}{11} \times 0{\cdot}2$$
$$= 8{\cdot}6 + 0{\cdot}1818$$
$$Q_1 = 8{\cdot}7818$$

The upper quartile $Q_3 = L + \frac{q_3 - c}{f} \times i$

$$\text{where } q_3 = \frac{3(85+1)}{4} = 64{\cdot}5$$

Hence the upper quartile class is 9·4—9·6 and so $L=9{\cdot}4$, $i=0{\cdot}2$, $f=10$, $c=57$

$$\therefore \quad Q_3 = 9{\cdot}4 + \frac{64{\cdot}5 - 57}{10} \times 0{\cdot}2$$
$$= 9{\cdot}4 + 0{\cdot}15$$
$$\therefore \quad Q_3 = 9{\cdot}55$$

ExampleNo. (10)

The classification of 75 cows has been done according to one day milk in the following table. Find the median ?

class interval (Milk in lbs)	No. of cows
8—10	4
10—12	8
12—14	12
14—16	25
16—18	15
18—20	8
20—22	3

Solution:—

We first calculate the cumulative frequencies which are given in the following table—

class interval (x)	frequency (f)	c.f.
8—10	4	4
10—12	8	12
12—14	12	24
14—16	25	49
16—18	15	64
18—20	8	72
20—22	3	75

The median is computed by the formula—

$$Md = L + \frac{m-c}{f} \times i$$

$$\text{where } m = \frac{n+1}{2} = \frac{75+1}{2} = 38$$

Hence the median lies in the class 14—16 and so $L = 14$, $f = 25$, $c = 24$, $i = 2$

$$Md = 14 + \frac{38-24}{25} \times 2$$

$$= 14 + 1{\cdot}12 = 15{\cdot}12$$

$\therefore$ Median $= 15{\cdot}12$

Example No. 11

The following table gives the marks obtained by a batch of candidates in a certain examination of History and politics. In which subject is the level of knowledge of the candidates higher ? Give reasons ?

Roll No.	History	Politics	Roll No.	History	Politics
1	42	46	9	40	30
2	24	20	10	62	61
3	38	41	11	55	50
4	35	43	12	54	63
5	30	25	13	52	45
6	45	54	14	47	56
7	58	47	15	43	58
8	50	36			(M.Sc. Agra 1955)

Solution:—

To find out the subject in which the knowledge of students is higher, we calculate the medians of the marks obtained in the two subjects. In the subject, for which the median is higher, the level of knowledge is higher.

For this, we arrange the marks in the ascending order of magnitudes in each subject.

History : 24, 30, 35, 38, 40, 42, 43, 45, 47, 50, 52, 54 ,55 58, 62

Politics : 20, 25, 30, 36, 41, 43, 45, 46, 47, 50, 54, 56, 58, 61, 63

The median will be the size of $\left\{\frac{n+1}{2}\right\}^{th}$ item

$=$size of 8^{th} item, since $n=15$

Median for History-marks$=45$

Median for Politics-marks$=46$

Hence, the level of knowledge of the candidates is higher in Politics than in the history.

Example No. (12)

Define Mode ? Illustrate its method of calculation ?

Solution:—

Mode : The most frequent or the most popular value of a variate is called the mode. It is repeated at the greatest no of times. In popular language, when we speak the average student or average rent, we generally imply the modal student or the modal rent. It is easy to locate, as it lies at the highest frequency. But if there are irregularities in the frequency distribution, the position of the mode could become indefinite. In such cases, the process of grouping will be applicable. In a discrete series the size of the variable which has the max. frequency is the mode; while in a continuous series the mode will be located by interpolation in the modal group by the formula

$$Mo=L+\frac{\triangle_1}{\triangle_1+\triangle_2}\times i$$

where L means the lower limit of the modal class

$\triangle_1$ ——→ Stands for the difference between the frequency of the modal class and that of the class which follows the modal class

$\triangle_2$ ——→ Stands for the difference between the frequency of the modal class and that of the the class which preceeds the modal class.

i ——→ Stands for the class interval of the modal class.

......

Let us consider the following age-distribution of the candidates appearing at the matriculation examination of Patna University in 1937—

Age-group (years)	No. of stud.	Age-group (years)	No. of stud.
12—13	5	17—18	980
13—14	48	18—19	981
14—15	189	19—20	794
15—16	303	20—21	515
16—17	522	21—22	474
Total			4811

Here the modal class is 18-19 corresponding to the highest frequency 981 and so,

$$L=18,\ i=19-18=1,\ \Delta_1=981-980=1,\ \Delta_2=981-794=187$$

$$\therefore \text{Mo}=18+\frac{1}{1+187}\times 1$$

$$=18+{\cdot}005=18{\cdot}005$$

$\therefore$ Mode$=18{\cdot}005$ years.

ExampleNo. (13)

Define dispersion ? What are the different methods of measuring it, describe each briefly ? (U.P. Board, 1962)

Solution:—

Having been known the value of the average of a series, we try to comput some statistics which can give an idea how the values of the variate are scattered around the central value. This variation of the data from the central value is called the *dispersion or Scatter or Spread or variability.*

Thus, the dispersion can be defined as the extent to which the magnitudes of the items differ from the central value.

The following are the measures of dispersion—

(i) Range

(ii) Quartile deviation

(iii) Mean deviation from (a) average (mean)
(b) median

(iv) Standard deviation

(i) Range : It is defined as the difference between the greatest and the least magnitude of the items. It is very easy to calculate but its use as a measure of dispersion is very rare. Because it does not depend upon all the observations (items) and in its computation no consideration is given to the value of the central tendency.

(ii, Quartile deviation : (Semi Inter quartile range) It is half of the quartile range i e.

$$Q.D.=\frac{Q_3-Q_1}{2}$$

It is a better measure of dispersion than the range and is oftenly used in elementary descriptive statistics. But due to its incapability of algebraic treatments, it is not used in the

Advance-theory. Another reason is that it gives an idea of dispersion only of those items which lie between Q_1 & Q_3 and so most of the items have no effect on the computation of dispersion.

(iii) Mean deviation (average deviatson) : As the computations of range & quartile deviation do not depend upon all the observation, so these measures of dispersion can not be said as the satisfactory measures of dispersion. One measure free from this objection is the mean deviation which is defined as the arithmatic average of the absolute deviations of the items from any measure (mean or median) of central tendency. Mathematically we write—

$$\text{Mean deviation} = \frac{\Sigma f|d|}{n},$$

where n is the total number of items,

f ⟶ Stands for the corresponding frequency of the variate x

and |d| ⟶ Stands for the absolute deviation of x from its average.

If the deviations are taken from mean, then it is called *Mean Deviation about mean* and if deviations are taken from the median, then it is called *Mean Deviation obout median*. The mean deviation also suffers from one drawback that is of incapability of further algebraic treatments.

For the illustration of the procedure, we calculate the mean deviation about mean from the following data—

x :	2·5,	7·5,	12·5,	17·5,	22·5,	27·5,	32·5
f :	4,	6,	10,	16,	12,	8,	4

To compute the mean deviation about the mean, we form the following table—

(1) x	(2) f	(3) f x	(4) d=x-M	(5) |d|	(6) f | d |
2·5	4	10·0	−15·5	15·5	62·0
7·5	6	45·0	−10·5	10 5	63 0
12·5	10	125·0	−5.5	5·5	55·0
17·5	16	280·0	− 0·5	0.5	8·0
22·5	12	270·0	+4·5	4·5	54·0
27 5	8	220 0	+9·5	9·5	76·0
32 5	4	130·0	+14·5	14·5	58·0
	60=Σf	1080 =Σf x			376·0 =Σf |d|

$$\text{Mean } \bar{X} \text{ or } M = \frac{\Sigma fx}{n} = \frac{1080}{60} = 18$$

First of all, we calculate the mean $\bar{X} = 18$, then we take the deviations of all values of the variate of col^{n}. (1) from their actual mean 18 as shown in col^{n}. (4), by 'd'. Again in col^{n}. (5), we write the absolute values of 'd' and in (6)th the product of (2) & (5) with respective figures.

$$\begin{array}{l}\text{Mean deviation}\\ \text{(about mean)}\end{array} = \frac{\Sigma f|d|}{n} = \frac{376}{60} = 6{\cdot}266$$

If we want to calculate the mean deviation about the median then first we find the median for the data and then tarke the deviations of the values from this median. Similarly, the mean deviation about any average can be found by first computing that desired average for the given data and then obtaining the deviations of the values from that average. The rest of the procedure is the same as indicated above.

(iv) Standard deviation : Among all the measures of dispersion, the S.D. is most widely used because of the facts that

(1) its computation is based on all the observations,

(2) the deviotions are taken from the mean and so the sum of squares of the deviations is minimum,

(3) it is capable of algebraic treatments.

The S D. is the square root of the mean of the squares of deviations from the arithmatic mean. Symbolically, the greek letter sigma (σ) denotes the standared deviation.

$$\sigma = \sqrt{\frac{\Sigma(x-\overline{X})^2}{n}} \quad \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

$$\text{and } \sigma = \sqrt{\frac{\Sigma f(X-\overline{X})^2}{n}} = \sqrt{\frac{\Sigma fx^2}{n} - \left(\frac{\Sigma fx}{n}\right)^2}$$

We can take the deviations from an assumed mean also inplace of the actual mean for our convenience. Then the formula for S. D. will be

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

and $\sigma = \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2}$ for a frequency distribution.

For illustration, we summarize the steps in the computation of S.D. by taking the following simple series of values—

X : 192, 288, 236, 229, 184, 230, 348, 291. 330, 243

(1) Choos the assumed mean 'A',

(2) Take the deviations 'd' from this assumed mean,

3. Square the deviations *i. e.* calculate 'd^2'.

The results of the above steps for the given series are tabulated below—

×	$d = X - A$ $A = 260$	d^2
192	-68	4624
288	$+28$	784
236	-24	576
229	-31	961
184	-76	5776
260	0	0
348	$+88$	7744
291	$+31$	961
330	$+70$	4900
243	-17	289
Totals	$+1 = \Sigma d$	26615 $= \Sigma d^2$

The S. D. is given by

$$\sigma = \sqrt{\left\{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2\right\}}$$

$$= \sqrt{\left\{\frac{26615}{10} - \left(\frac{1}{10}\right)^2\right\}}$$

$$= \sqrt{(2661{\cdot}5 - {\cdot}01)}$$

$$= \sqrt{(2661{\cdot}49)}$$

$$\therefore \quad \sigma \approx 51{\cdot}59.$$

To complete the S. D. in the case of frequency distribution, the following steps are made—

(1) Choose the assumed mean, 'A',

(2) take the deviations $d = (X - A)$ from this assumed mean,

(3) calculate the squares of d,

(4) compute the prroducts (fd) in a column,

(5) compute the products (fd^2) in another column.

For illustration of the procedure, we calculate the S. D. from the following data—

Height in cms. (X) : 60, 61, 62, 63, 64, 65, 66, 67, 68.

No. of plants (f) ; 2, 0, 15, 29, 25, 12, 10, 4, 3.

The computational work is done in the following tabular form—

$$A=64$$

X	f	$d=X-A$	d^2	fd	fd^2
60	2	-4	16	-8	32
61	0	-3	9	0	0
62	15	-2	4	-30	60
63	29	-1	1	-29	29
64	25	0	0	0	0
65	12	1	1	$+12$	12
66	10	2	4	$+20$	40
67	4	3	9	$+12$	37
68	3	4	16	$+12$	48
Totals	100			$-11=\Sigma fd$	$257=\Sigma fd^2$

The S. D. is given by—

$$\sigma=\sqrt{\left\{\frac{\Sigma fd^2}{n}-\left(\frac{\Sigma fd}{n}\right)^2\right\}}=\sqrt{\left\{\frac{257}{100}-\left(\frac{-11}{100}\right)^2\right\}}=\sqrt{(2{\cdot}57-{\cdot}0121)}$$

$$=\sqrt{(2{\cdot}5579)}$$

$$\approx 1{\cdot}59$$

Computation of the S. D. in the case of frequency distribution of the grouped data with unequal Interval.

In this case, the classes are replaced by their mid-values and the rest procedure remains the same as explained above.

In the case of a grouped data when the class-intervals are equal, a more convenient formula for calculating the S. D. is.

$$\sigma=i\times\sqrt{\left\{\frac{\Sigma f\,\xi^2}{n}-\left(\frac{\Sigma f\,\xi}{n}\right)^2\right\}},$$ where i is the class-interval and

$$\xi=\frac{d}{i}$$

This is known as the *method of step-deviation.*

The steps according to this method are given below—

(1) choose the assmed mean 'A',

(2) take the deviations $d=(X-A)$ from this assumed mean,

(3) divide the deviations by the class interval say i and call the value $\frac{d}{i}=\xi$

(4) calculate the squares of ξ

(5) compute the products $f\xi$, and

(6) finally compute the products $f\xi^2$.

For illustration, consider the following example—

Age group (X) : 20—30, 30—40, 40—50, 50—60, 60—70,
(Years) 70—80, 80—90.
No. of persons (f) : 3, 61, 132, 153, 140, 51, 2

The computational work for S. D. is shown in the following tabular form—

$A=55, \qquad i=10$

Age group (Years)	Mid value X	f	$d=(X-A)$	$\xi=d/i$	$f\xi$	$f\xi^2$
20—30	25	3	−30	−3	−9	27
30—40	35	61	−20	−2	−122	244
40—50	45	132	−10	−1	−132	132
50—60	55	153	0	0	0	0
60—70	65	140	+10	+1	+140	140
70—80	75	51	+20	+2	+102	204
80—90	85	2	+30	+3	+6	18
Totals		$542=\Sigma f$			$-15=\Sigma f\xi$	$765=\Sigma f\xi^2$

The S. D. is given by

$$\sigma=i\times\sqrt{\left\{\frac{\Sigma f\xi^2}{n}-\left(\frac{\Sigma f\xi}{n}\right)^2\right\}}$$

$$=10\times\sqrt{\left\{\frac{765}{542}-\left(\frac{-15}{542}\right)^2\right\}}$$

$$=10\times\sqrt{(1\cdot4114-\cdot0008)}=10\times\sqrt{(1\cdot4106)}$$
$$=10\times1\cdot187$$
$$\approx11\cdot87$$

$\therefore$ $\sigma=11\cdot87$ approximately.

Exp. 14. Prove that $\Sigma x^2=\Sigma X^2-\frac{(\Sigma X)^2}{n}$

Where x is the deviation of an observation $\bar{x}$ from the arithmatic mean $\bar{X}$ of the sample and n is the total number of observations in the sample.

(M. Sc. Ag. Agra, 1963)

Sol. We have $x=X-\bar{X}$, where $\bar{X}=\frac{\Sigma X}{n}$

$$\therefore \quad x^2=(X-\bar{X})^2$$
$$\text{and } \Sigma x^2=\Sigma(X-\bar{X})^2$$
$$=\Sigma[X^2+\bar{X}^2-2X\bar{X}^2]$$
$$=\Sigma X^2+\Sigma\bar{X}^2-2\bar{X}\Sigma X$$
$$=\Sigma X^2+n\bar{X}^2-2n\bar{X}^2 \qquad \because \Sigma X=n\bar{X}$$
$$=\Sigma X^2-n\bar{X}^2$$
$$=\Sigma X^2-n.\left(\frac{\Sigma X}{n}\right)^2$$
$$\therefore \quad \Sigma x^2=\Sigma X^2-\left(\frac{\Sigma X}{n}\right)^2$$

Hence Proved.

Exp. 15. Compute the S. D. of the following data by means of the formulae—

(a) $\sigma^2=\frac{\Sigma f(X-\bar{X})^2}{n}$

(b) $\sigma^2=\frac{\Sigma fX^2}{n}-\left(\frac{\Sigma fX}{n}\right)^2$

and (c) $\sigma^2=\frac{\Sigma fd^2}{n}-\left(\frac{\Sigma fd}{n}\right)^2$

x :	12,	13,	14,	15,	16,	17,	18,	20
f :	4,	11,	32,	21,	15,	8,	5,	4

Sol. : (a) For the computation of the S. D., the table is given below—

$\bar{X}=15$

X	f	fX	$d=(X-\bar{X})$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
12	4	48	−3	9	36
13	11	143	−2	4	44
14	32	448	−1	1	32
15	21	315	0	0	0
16	15	240	1	1	15
17	8	136	2	4	32
18	5	90	3	9	45
20	4	80	5	25	100
Totals	$\Sigma f=100$	$\Sigma fX=1500$			$304X=$ $\Sigma f(X- \ )$

$$\bar{X}=\frac{\Sigma fX}{\Sigma f}=\frac{1500}{100}=15$$

Thus the S. D. is given by

$$\sigma=\sqrt{\left\{\frac{\Sigma f(X-\bar{X})^2}{n}\right\}}=\sqrt{\frac{304}{100}}=\sqrt{(3{\cdot}04)}$$

∴ $\sigma=1{\cdot}74$ approximately.

(b) To compute the S. D., the following table is framed—

X	f	X^2	fX	fX^2
12	4	144	48	576
13	11	169	143	1859
14	32	196	448	6272
15	21	225	315	4725
16	15	256	240	3840
17	8	289	136	2312
18	5	324	90	1620
20	4	400	80	1600
Totals	$100=\Sigma f$		$\Sigma fX=1500$	$22{,}804=\Sigma fX^2$

The S. D. is given by

$$\sigma=\sqrt{\left\{\frac{\Sigma fX^2}{n}-\left(\frac{\Sigma fX}{n}\right)\right\}}=\sqrt{\left\{\frac{22{,}804}{100}-\left(\frac{1500}{100}\right)^2\right\}}$$

$$=\sqrt{(228{\cdot}04-225}$$

$$\therefore \qquad \sigma=\sqrt{3{\cdot}04}=1{\cdot}76 \text{ approximately.}$$

(c) The S. D. is computed with the help of the following table—

$A=16$

X	f	$d=(X-A)$	d^2	fd	fd^2
12	4	−4	16	−16	64
13	11	−3	9	−33	99
14	32	−2	4	−64	128
15	21	−1	1	−21	21
16	15	0	0	0	0
17	8	1	1	8	8
18	5	2	4	10	20
20	4	4	16	16	64
Totals	$100=\Sigma f$			$-100=\Sigma fd$	$404=\Sigma fd^2$

$$\sigma=\sqrt{\left\{\frac{\Sigma fd^2}{n}-\left(\frac{\Sigma fd}{n}\right)^2\right\}}=\sqrt{\left\{\frac{404}{100}-\left(\frac{-100}{100}\right)^2\right\}}$$

$$=\sqrt{(4{\cdot}04-1)}$$

$$=\sqrt{3{\cdot}04}$$

$$\therefore \qquad \sigma=1{\cdot}76 \text{ approximately.}$$

The calculations have been considerabely reduced by using $\sigma=\sqrt{\left\{\frac{\Sigma fd^2}{n}-\left(\frac{\Sigma fd}{n}\right)^2\right\}}$ and so it is called *short cut formula.*

Exp. 16 Calculate the median and the quartiles using
(a) mathematical method, and
(b) graphical method, of the following data—

Class	Frequency
5—7	4
7—9	6
9—11	10
11—13	12
13—15	6
15—17	5
17—19	4

Sol. : First we prepare the cumulative frequency table—

Class	frepuency	*c. f.*
5—7	4	4
7—9	6	10
9—11	10	20
11—13	12	32
13—15	6	38
15—17	5	43
17—19	4	$47 = n$

We have

$$md = \text{L} + \frac{m-c}{f} \times i$$

where $m = \frac{n+1}{2} = 24$

So, median lies in the class 11—13

$$\text{M}d = 11 + \frac{24-20}{12} \times 2$$
$$= 11 + 0{\cdot}66$$

$\therefore$ Median $= 11{\cdot}66$

The lower quartile is,

$\text{Q}_1 = \text{L} + \frac{q_1 - c}{f} \times i$ where $q_1 = \frac{n+1}{1} = \frac{47+1}{4} = 12$ and so,

$$= 9 + \frac{12-10}{10} \times 2$$

$$=a^2\,\frac{\Sigma x^2}{n}+2ab\,\frac{\Sigma xy}{n}+b^2\,\frac{\Sigma y^2}{n}-\left\{a^2\left(\frac{\Sigma x}{n}\right)^2\right.$$

$$\left.+b^2\left(\frac{\Sigma y}{n}\right)^2+2ab\,\frac{\Sigma x}{n}\,\frac{\Sigma y}{n}\right\}$$

or
$$\sigma_z{}^2=a^2\left\{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2\right\}+b^2\left\{\frac{\Sigma y^2}{n}-\left(\frac{\Sigma y}{n}\right)\right\}$$

$$+2ab\left\{\frac{\Sigma xy}{n}-\bar{x}\,.\,\bar{y}\right\}$$

$$=a^2\,.\,\sigma_x{}^2+b^2\sigma_y{}^2+\frac{2ab}{n}\,\{\Sigma xy-n\bar{x}\bar{y}\}$$

$$=a^2\;\sigma_x{}^2+b^2\sigma_y{}^2+\frac{2ab}{n}\;\Sigma\,(x-\bar{x})\;(y-\bar{y})$$

$\therefore\;\; \sigma_z{}^2=a^2\sigma_x{}^2+b^2\sigma_y{}^2+0$

Because $\Sigma\,(x-\bar{x})\;(y-\bar{y})=0$ as x and y are independent variates.

$\therefore\;\; \sigma_z{}^2=a^2\sigma_x{}^2+b^2\sigma_y{}^2.$ *Hence proved.*

(c) $V\,(x+y)=V\,(x)+V\,(y)$ when x and y are indep. and $a=1$, $b=1$

$=0{\cdot}5+1{\cdot}2$

$=1{\cdot}7.$

$\therefore$ S. D. $(x+y)=\sqrt{(1{\cdot}7)}$

or $\sigma_{(x+y)}=\mathbf{1{\cdot}3}$ approximately.

Similarly, $V\,(x-y)=V\,(x)+V\,(y)$, where $a=1$, $b=-1$

$=0{\cdot}5+1{\cdot}2$

$=1{\cdot}7$

$\therefore$ S. D. $(x-y)=\sqrt{(1{\cdot}7)}$

$\sigma_{(x-y)}=\mathbf{1{\cdot}3}$ approximately.

Also, $V\,(3x-4y)=9\times V\,(x)+16\times V\,(y)$

$=9\times0{\cdot}5+16\times1{\cdot}2$ as $a=3$,

$=4{\cdot}5+19{\cdot}2$ $b=-4$

$=23{\cdot}7$

S. D. $(3x-4y)=\sqrt{(23{\cdot}7)}$

or $\sigma_{(3x-4y)}=\mathbf{4{\cdot}86}$ approximately.

Exp. 18. Write notes on :

(a) Variance,

(b) Standard error,

(c) Coefficient of variation.

Sol. (a) **Variance :** It is the square of the standard deviation i.e.

$$\text{Variance}=(\text{S. D.})^2=\frac{\Sigma(X-\overline{X})^2}{n}$$

where X →represents the variate values

$\overline{X}$ →the mean of the variate values

n →the total number of items.

Thus, variance is the arithmatic mean of the squares of the deviations from the mean. It is denoted by σ^2 *and is widely used in the statistical analysis of field experiments.*

(b) **Standard error :** It is the standard deviation of any statistic calculated on the basis of sample observations. It is widely used in the testing of statistical hypotheses. The S. E. of the mean is given by $\frac{\text{S. D.}}{\sqrt{n}}$, where n is the total number of observations, on which the mean has been calculated i.e.

$$\text{S. E.}=\frac{\sigma}{\sqrt{n}}\cdot$$

(c) **Coefficient of variation :** The S. D. and other measures of dispersion are the absolute measures of dispersion and are expressed in the same units in which the observations are given and hence cannot be used to compare the variations in the two given series which differ in their units and averages. To compare the variations of two such series, we require some relative measure of dispersion. *The coefficients of variation is such a measure which is defined as the ratio of the S. D. to the mean expressed in percentage.* Symbolically,

$$\text{C. V.}=\frac{\sigma}{M}\times 100.$$

This statistic was developed by **Pearson**. In order to compare the variations in the two series, we compute the coefficients of variation for each of them. The series for which the C. V. is higher, will vary more than the other.

Exp. 19. The scores of two golfers for 10 rounds each are :

A : *58, 59, 60, 54, 65, 66, 52, 75, 69, 52*
B : *84, 56, 92, 65, 86, 78, 44, 54, 73, 68*

which may be regarded as the more consistent player ?

Sol. In the present problem, we have to compute the coefficients of variation in the two cases and then compare them. The one, which has less variability than the other, will be called as the more consistent player ?

The computation is made in the following tabular form :

Rounds	Golfer A			Golfer B		
	Scores (X)	$(X-\overline{X})$	$(X-\overline{X})^2$	Scores (Y)	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$
1	58	-3	9	84	$+14$	196
2	59	-2	4	56	-14	196
3	60	-1	1	92	$+22$	484
4	54	-7	49	65	-5	25
5	65	$+4$	16	86	$+16$	256
6	66	$+5$	25	78	$+8$	64
7	52	-9	81	44	-26	676
8	75	$+14$	196	54	-16	256
9	69	$+8$	64	73	$+3$	9
10	52	-9	81	68	-2	4
Totals	$610=\Sigma X$	0	$526=\Sigma(X-\overline{X})^2$	$700=\Sigma Y$	0	$\Sigma(Y-\overline{Y})^2=2166$

$$\overline{X}=\frac{\Sigma X}{n}=\frac{610}{10}=61 \qquad \overline{Y}=\frac{\Sigma Y}{n}=\frac{700}{10}=70$$

$$\text{S. D. } (\sigma_x)=\sqrt{\left(\frac{\Sigma(X-\overline{X})^2}{n}\right)} \qquad \text{S. D. } (\sigma_y)=\sqrt{\left(\frac{\Sigma(Y-\overline{Y})^2}{n}\right)}$$

$$=\sqrt{\left(\frac{526}{10}\right)} \qquad =\sqrt{\left(\frac{2166}{10}\right)}$$

$=\sqrt{(52{\cdot}6)}$

$=7{\cdot}252$ approx.

$$\text{C. V.}_x=\frac{\sigma}{M}\times 100=\frac{7{\cdot}252}{61}\times 100$$

$=11{\cdot}88$

$=\sqrt{(216{\cdot}6)}$

$=14{\cdot}717$ approx.

$$\text{C. V.}_y=\frac{14{\cdot}717}{70}\times 100$$

$=21{\cdot}024$

Sol. Comparing the two C. Vs., we conclude that the player A is more consistant than player B.

Exp. 20. (a) Explain the practical utility of the coefficient of variation in the Biological research ?

(b) The average heights of ten years old and 18 years old girls were reported to be 74·4 and 161·0 cms. respectively with standard deviations 2·64 and 6·12 cms. respectively. Which data varies more ? *(M. Sc. Ag. Agra 1963)*

Sol. (a) It has been answered in the above question.

(b) C. V. for the heights of 10 years old girls is

$$\text{C. V.}_1=\frac{2{\cdot}64}{74{\cdot}4}\times 100$$
$$=3{\cdot}548.$$

C. V. for the heights of 18 years old girls is

$$\text{C. V.}_2=\frac{6{\cdot}12}{161{\cdot}0}\times 100$$
$$=3{\cdot}801.$$

Comparing the two C. Vs., we conclude that the heights of the 18 years old girls show the greater variability than that of the 10 years old girls.

Exp. 21. A random sample from a biological population is given below :

Observation No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of grains/ ear head	19	30	11	31	32	36	39	24	12	57	39	29	34	53	33

In the above series, sum of all the observations is 479 and sum of squares of deviations from the mean is 2272·93.

(a) Calculate the following of the above sample.

Mean, variance, standard deviation and standard error.

(b) What is an array ? Illustrate your answer with the series en above.

(c) In samples of 15–20 items, the range is on the average about 3·5 times the standard deviation. Verify this statement from the sample given above ? *(M. Sc. Ag. Agra, 1964)*

Sol. (a) We have

$$n=15$$

$$\Sigma X=479$$

$$\Sigma (X-\bar{X})^2=2272\cdot93.$$

Hence, the mean $(\bar{X})=\frac{\Sigma X}{n}=\frac{479}{15}=31\cdot933$ grains/ear head

$$\text{Variance } (\sigma_x{}^2)=\frac{\Sigma(X-\bar{X})^2}{n}=\frac{2272\cdot93}{15}=151\cdot5286$$

$$\text{S. D. } (\sigma_x)=\sqrt{\left(\frac{\Sigma(X-\bar{X})^2}{n}\right)}=\sqrt{\left(\frac{2272\cdot93}{15}\right)}$$

$$=\sqrt{(151\cdot5286)}=12\cdot309$$

$$\text{S. E. (of mean)}=\frac{\text{S. D.}}{\sqrt{n}}=\sqrt{\left(\frac{\text{variance}}{n}\right)}=\sqrt{\left(\frac{151\cdot5286}{15}\right)}$$

$$=\sqrt{(10\cdot1019)}$$

$$=3\cdot17 \text{ approx.}$$

(b) **Array :** An orderly arrangement of the variate values is called an array. If the variate values i.e. the number of grains/earhead are arranged either in ascending or descending order of their magnitudes, the resulting series will be called an array. The practical utility of an array lies in the calculation of partition values.

(c) In the given sample, the lowest number of grains per earhead is 11 and highest is 57, so the range will be given by :

$$R=57-11=46 \text{ grains/earhead}$$

We have the S. D.$=12\cdot309$

so $3\cdot5\times\text{S. D.}=3\cdot5\times12\cdot309$

$$=43\cdot08$$

which is nearly equal to the range value i.e. 46 grains/earhead.

Thus, we can say that on the average the range is 3·5 times that of the S. D. in the sample given above.

Exp. 22. A random sample from a field experiment is given below :

Obs. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Hts. of Plants in cms.	23	17	20	29	18	22	16	25	13	15	19	21	23	20	21	22	16

In this sample, the sum of all the observations is 340 and the sum of squares of deviations from the mean is 254.

(a) Calculate six statistical constants, three each of measures of type and measures of variability, from the above sample.

(b) Explain what is an array, with the help of the above sample. What is its practical utility ? *(M. Sc. Ag. Agra, 1965)*

Sol. (a) We have,

$$n=17$$

$$\Sigma X=340$$

$$\Sigma (X-\bar{X})^2=254$$

The three measures of central tendency are—Mean, Median, Harmonic mean (as the Mode cannot be found in the present example without the frequencies).

Then Mean $(\bar{X})=\frac{\Sigma X}{n}=\frac{340}{17}=20{\cdot}0$ cms.

The median can be found out by arranging the values (heights of plants in cms) in ascending order of their magnitudes. The arrangement will be as follows :

13, 15, 16, 16, 17, 18, 19, 20, 20, 21, 21, 22, 22, 23, 23, 25, 29

$$\text{Md.}=\text{size of the } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item}=\text{size of } \left(\frac{17+1}{2}\right)^{\text{th}} \text{ item}$$

$$=\text{size of 9th item.}$$

The size of the 9th item in the orderly arranged series is 20 and so the

$$\text{Median}=20{\cdot}0 \text{ cms.}$$

For the harmonic mean, we know that

$$\text{H. M.} = \frac{n}{\Sigma \frac{1}{X}}$$

and so it is computed with the help of the following table :

S.N.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Totals
X	23	17	20	29	18	22	16	25	13	15	19	21	23	20	21	22	16	340
$\frac{1}{X}$	·0434	·0588	·0500	·0344	·0555	·0454	·0625	·0400	·0770	·0666	·0526	·0450	·0434	·0500	·0450	·0454	·0625	0·8575

$$\therefore \quad \text{H. M.} = \frac{17}{0{\cdot}8575} = 19{\cdot}9 \text{ approximately.}$$

The three averages (measures of central tendency) are found to be equal.

For the three measures of dispersion, we take the range, the S. D. and the mean deviation (about the mean) as computed below :

$$\text{Range (R)} = \text{highest value} - \text{lowest value}$$

$$= 29 - 13 = 16 \text{ cms.}$$

$$\text{S. D. } (\sigma_x) = \sqrt{\left(\frac{\Sigma(X-\bar{X})^2}{n}\right)} = \sqrt{\left(\frac{254}{17}\right)} = \sqrt{(14 \cdot 942353)}$$

$$= 3 \cdot 865 \text{ cms. approximately.}$$

$$\textbf{Mean Deviation (M. D.)} = \frac{\Sigma \mid d \mid}{n},$$

where $|d|$ means the absolute value of the deviation from the actual mean $\bar{X}$ i.e. $|d| = |X - \bar{X}|$.

To compute it, we have to frame the following table :

We have $\bar{X}=20{\cdot}0$ cms.

S. N.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Totals
x	23	17	20	29	18	22	16	25	13	15	19	21	23	20	21	22	16	340
$d=X-\bar{X}$	3	−3	0	9	−2	2	−4	5	−7	−5	−1	1	3	0	1	2	−4	0
$\mid d \mid$	3	3	0	9	2	2	4	5	7	5	1	1	3	0	1	2	4	52

$\therefore$ Mean deviation $=\frac{52}{17}=3{\cdot}0588$ cms. approx.

(about mean)

Hence we have, the three calculated—

Measures of average	*Measures of dispersion*
Mean $=20{\cdot}0$ cms.	Range$=16$ cms.
Median$=20{\cdot}0$ cms,	S. D.$=3{\cdot}865$ cms.
H. M. $=19{\cdot}9$ cms.	M. D.$=3{\cdot}0588$ cms.

(b) The answer has been given in the previous question.

Exp. 23. Account of the number of grains on each of fourty earheads of wheat gave the following results—

32	23	25	19	26	17	31	23
29	39	34	26	27	38	27	34
33	27	29	27	31	38	37	39
18	34	35	40	34	43	24	31
28	43	20	33	28	34	29	28

Present the data in the form of a frequency distribution. Calculate the mean and standard deviation of the distribution from the grouped data and state the standard error of mean number of grains per earhead, *(M. Sc. Ag. Agra 1958)*

Sol. The lowest magnitude is 17 and highest is 43. Taking the lower limit of the 1st class 16, and keeping the class-internal equal to 4, we shall obtain the following seven classes with their respective frequencies. *The border lines items are placed in the classes where they are as the lower limits.*

Class	Tally marks	Frequency
16—20	|||	3
20—24	||||	4
34—28	𝍸 ||	7
28—32	𝍸 |||	8
32—36	𝍸 𝍸	10
36—40	𝍸	5
40—44	|||	3

To compute the mean, S. D. and the standard error af mean number of grains per earhead, we form the following table—

$A=30$

Mid value X	$d=X-A$	d^2	f	fd	fd^2
18	-12	144	3	-36	432
22	-8	64	4	-32	256
26	-4	16	7	-28	112
30	0	0	8	0	0
34	4	16	10	40	160
38	8	64	5	40	320
42	12	144	3	36	432
Totals	—	—	$40=\Sigma f$	$+20$ $=\Sigma fd$	1712 $=\Sigma fd$

Mean is given by

$$\begin{aligned}\bar{X}&=A+\frac{\Sigma fd}{n}\\&=30+\frac{20}{40}\\&=30+0{\cdot}5\\\bar{X}&=\mathbf{30{\cdot}5}\text{ grains}\end{aligned}$$

$$\text{S. D. }(\sigma)=\sqrt{\left\{\frac{\Sigma fd^2}{n}-\left(\frac{\Sigma fd}{n}\right)^2\right\}}=\sqrt{\left\{\frac{1712}{40}-\left(\frac{20}{40}\right)^2\right\}}$$

$$=\sqrt{(42{\cdot}80-0{\cdot}25)}\sqrt{(42{\cdot}55)}.$$

$\therefore$ S. D. = **6·52** approximately.

S. E. of mean no. of grains/earhead is given by

$$\text{S. E.}=\frac{\text{S. D.}}{\sqrt{n}}=\frac{6{\cdot}52}{\sqrt{(40)}}=\frac{6{\cdot}52}{6{\cdot}32}=1{\cdot}03.$$

$\therefore$ Standard error = **1·03** grains.
Mean = 30·5 grains
S. D. = 6·52 grains
S. E. of mean no. of grains/earhead
= 1·03 grains ... ... **Ans.**

Exp. 24. The length of 50 earheads of wheat are given below in centimetres—

10·4	10·8	8·5	10·9	10·5	9·4	11·1	7·0	8·8	10·6
10·9	11·2	10·5	9·7	10·8	8·4	9·3	11·2	7·0	11·3
12·2	8·9	9·6	11·0	10·4	9·9	9·0	10·5	9·6	11·0
11·5	9·7	10·6	10·4	9·5	8·4	8·7	9·1	11·5	10·6
11·8	7·8	10·1	10·1	8·4	11·0	9·5	9·4	9·8	11·7

(a) Present the data in the form of a frequency distribution by choosing a suitable class-interval.

(b) Calculate the mean and median of the distribution and represent it graphically. *(M. Sc. Agra, 1960)*

Sol. (a) The lowest magnitude is 7·0 and highest is 12·2. Taking the lower limit of the first class 7·0 and keeping the class-interval equal to 0·8, we obtain the following seven classes with their respective frequencies written against them. *The border line items are placed in the classes where they are as the lower limits*—

Class	Tally marks	Frequency
7·0—7·8	||	2
7·8—8·6	𝍸	5
8·6—9·4	𝍸 |	6
9·4—10·2	𝍸 𝍸 |||	13
10·2—11·0	𝍸 𝍸 |||	13
11·0—11·8	𝍸 ||||	9
11·8—12·6	||	2
Totals	50	50

(b) To calculate the mean of the distribution, we form the following table—

Mid Value X	$d=X-A$	f	fd
7·4	−2·4	2	−4.8
8·2	−1·6	5	−8·0
9·0	−0·8	6	−4·8
9·8	0	13	0
10·6	0·8	13	10·4
11·4	1·6	9	14·4
12·2	2·4	2	4·8
Totals	—	$\Sigma f=50$	12·0 $=\Sigma fd$

Hence the mean is

$$\bar{X}=A+\frac{\Sigma fd}{n}$$

$$=9{\cdot}8+\frac{12{\cdot}0}{50}$$

$$=9{\cdot}8+0{\cdot}24$$

$$\therefore \quad \bar{X}=10{\cdot}04 \text{ centimetres}$$

For the computation of median of the distribution we prepare the following table—

Class	freq. (f)	c. f
7·0—7·8	2	2
7·8—8·6	5	7
8·6—9·4	6	13
9·4—10·2	13	26
10·2—11·0	13	39
11·0—11·8	9	48
11·8—12·6	2	50

Thus the median is

$$Md=L+\frac{m-c}{f}\times i$$

where $m=\frac{n+1}{2}=\frac{50+1}{2}=25{\cdot}5$

Hence the median lies in the class 9·4—10·2 and so

$L=9{\cdot}4,\ c=13,\ f=13,\ i=0{\cdot}8$

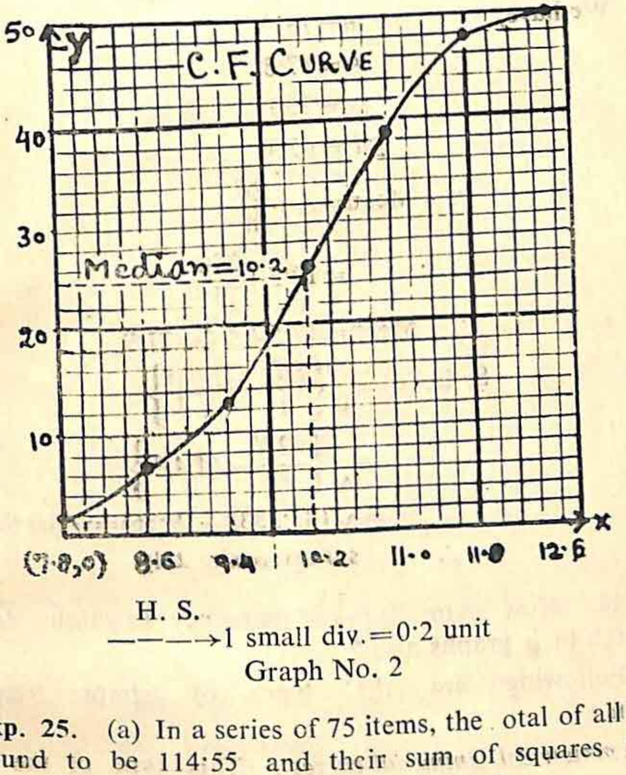
$$\therefore \quad Md = 9{\cdot}4 + \frac{25{\cdot}5 - 13}{13} \times 0{\cdot}8$$

$$= 9{\cdot}4 + 0{\cdot}7692$$

$$= 10{\cdot}1692$$

$$\therefore \quad Md \approx 10{\cdot}17 \text{ centimetres.}$$

Graphically the median = 10·2 cms. and it is shown in the adjacent graph.



V. S. ——→1 small div. = 2 units

H. S.
——→1 small div. = 0·2 unit
Graph No. 2

Exp. 25. (a) In a series of 75 items, the total of all the items was found to be 114·55 and their sum of squares 175·7125. Calculate the mean and the standard deviation ?

(b) For the marks obtained by 75 students in a certain test, the sum of the deviation from the assumed mean 17·5 was 330 and the sum of squares of the deviations was 6250. Calculate the mean and the standard deviation ?

Sol. (a) We are given $n = 75$

$$\Sigma x = 114{\cdot}55$$

$$\Sigma x^2 = 175{\cdot}7125.$$

$$\therefore \quad \textbf{Mean} = \frac{\Sigma x}{n} = \frac{114{\cdot}55}{75} = \mathbf{1{\cdot}527}$$

$$\text{S. D. } (\sigma)=\sqrt{\left\{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2\right\}}$$

$$=\sqrt{\left\{\frac{175\cdot7125}{75}-(1\cdot527)^2\right\}}$$

$$=\sqrt{(2\cdot3429-2\cdot3327)}$$

$$=\sqrt{(0\cdot0102)}.$$

$\therefore$ $\sigma=\mathbf{0\cdot101}$ approximately.

(b) We have,

$$n=75$$

$$A=17\cdot5$$

$$\Sigma d=330$$

$$\Sigma d^2=6250$$

$$\therefore \quad \text{Mean}=A+\frac{\Sigma d}{n}$$

$$=17\cdot5+\frac{330}{75}$$

$$\therefore \quad \text{Mean}=17\cdot5+4\cdot4=\mathbf{21\cdot9}.$$

$$\text{S. D. } (\sigma)=\sqrt{\left\{\frac{\Sigma d^2}{n}-\left(\frac{\Sigma d}{n}\right)^2\right\}}$$

$$=\sqrt{\left\{\frac{6250}{75}-(4\cdot4)^2\right\}}$$

$$=\sqrt{(83\cdot3333-19\cdot36)}=\sqrt{(63\cdot9733)}$$

$\therefore$ $\sigma=\mathbf{8\cdot0}$ approximately.

Exp. 26. How many types of frequency functions do you know ; sketch their graphs also ?

Sol. Following are the types of simple frequency distributions :

(i) *Symmetrical Unimodal curves :* This type of the curve is symmetrical about its maximum ordinate. The class-frequencies

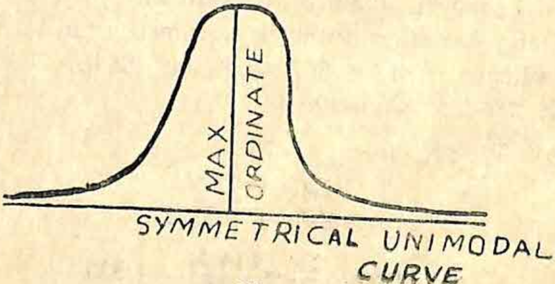


Fig. 1

go on decreasing symmetrically to zero on both the sides of the central (maximum) ordinate. A very important of this type is the normal curve.

(2) *Skewed Unimodal curves :* In this type of curves, the frequencies fall more rapidly on one side of the maximum ordinate than that of the other. Such curves have a tail on one side of the max. ordinate. If the tail is on the right hand side of the max. ordinate,the curve is siad to be the *+vely Skewed.*

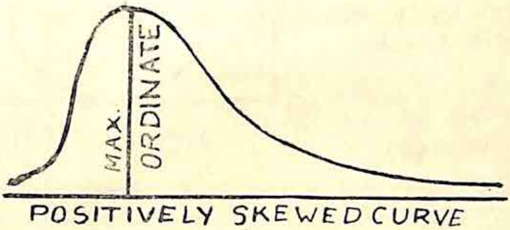


Fig. 2 (a)

When the tail is on the left hand side of the max. ordinate the curve is said *—vely Skewed curves.*

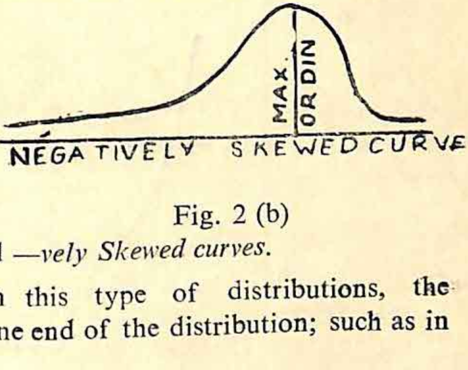


Fig. 2 (b)

(3) *J-Shaped curves :* In this type of distributions, the maximum frequency falls on one end of the distribution; such as in the bank-balances and other income-distributions. They may be +vely. or —vely J-shaped.

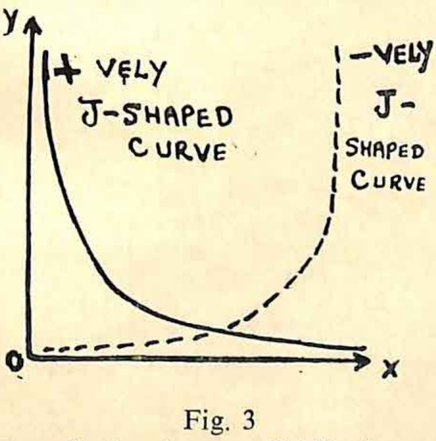


Fig. 3

(4) *U-shaped or anti-modal curves :* In this type of curves, the frequencies start from a maximum value and then fall to a minimum and again increase. These curves may be symmetrical or may not be. Such type of distributions occur in the number of unemployed persons by age-groups.

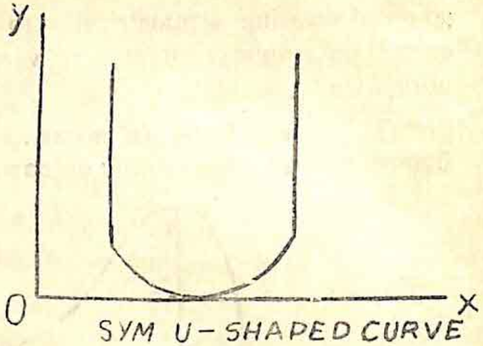


Fig. 4 (a)

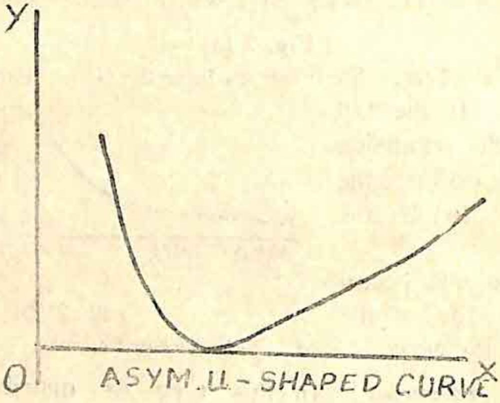


Fig. 4 (b)

EXERCISE No. (I)

(Problems on Ch. I, II, III)

Q. 1. The yield of rice in 55 equal plots of a village is given in maunds as follows—

62 59 23 27 77 33 43 13 47 89 32
63 24 17 72 30 45 42 8 26 36 21
25 13 64 37 12 48 5 29 41 40 51
35 67 36 18 28 92 92 2 20 55 27
65 52 15 21 78 34 40 16 57 81 31 Prepare a frequency .table taking a suitable class-interval and draw the histogram, frequency polygon and the frequency curve.

(U. P. Board, 1964)

Q. 2. (a) What is a frequency distribution? What considerations are involved in forming a frequency table from a body of observational data on a quantitative character?

(M. Sc. Ag. Agra, 1956)

(b) Make a freqcency table having grades of wages with class-intervals of five annas each from the following data of daily wages received by 40 labourers in a certain factory—

5, 15, 30, 22, 30, 25, 40, 10
6 20 15 25 32 22 20 10
7 8 17 20 32 22 11 15
8 11 11 22 8 35 25 35
11 25 6 15 20 37 20 42

Q. 3. Draw the ogive of the classified data obtained from Q. No. (2) (b) and find out median & quartilies from it, graphically?

Q. 3. In the following table, the heights of 2 plants are given in centimeteres at various ages—

Age (days)	Height of plant	
	A in cm.	B in cm.
40	153	156
60	167	173
80	182	180
100	187	185
120	198	195
140	201	200
160	219	219
180	220	231
200	220	231

Draw the graphs for the growth of the two plants and comment on their growth.

Q. 5. (a) Represent the following data by a suitable diagram—

Dose of nitrogen in Kgm./Hectre	Height of plants in cms. after	
	30 days	70 days
0	28	52
22	35	64
44	40	75
66	38	67
70	37	63

(b) The following table gives the total outlay on rural development proposed in the first five years plan and its break downinto major items. Give a suitable diagram to represent the data—

Item	Amount (in crores of rupees)
Agr. & community development	371·43
Irrigation	178·97
Irrigation and power	276·90
Power	138·54
Transport and communication	508·10
Industry	184·04
Social Services	350·81
Rehabilitation	96·00
Miscellaneous	62·99
Total	2167·78

Q. 6. The consumer's price-index and its three most important component indexes for a country from 1935 to 1944 are given in the table below :

Year	Consumer's Price Index	Food	Clothing	Rent
1935	98·1	100·4	96·8	94·2
1936	99·1	101·3	97·6	96·4
1937	102·7	105·3	102·8	100·9
1938	100·8	97·8	102·2	104·1
1939	99·4	95·2	100·5	104·3
1940	100·2	96·6	101·7	104·6
1941	105·2	105·5	106·3	106·2
1942	116·5	123·9	124·2	108·5
1943	123·6	138·0	129·7	108·0
1944	125·5	136·1	138·8	108·2

Using graphical-method, comment particularly in respect of the periods 1937–38, 1939–40 and 1941–44.

(M. Sc. Ag. Eco. Agra, 1965)

Q. 7. The following data gives the marks obtained by a batch of candidates in a certain examination in Mathematics and Statistics. In which subject is the level of knowledge of the candidates higher ? Give reasons ?

Marks in Maths. :	14, 16, 16, 14, 22, 13, 15, 24, 12, 23, 14, 20, 17
Marks in Stat. :	21, 22, 18, 18, 19, 20, 17, 16, 15, 11, 12, 21, 20

Q. 8. (a) Define the median and its utility ? Calculate the median from the following data :

Class Interval	9–25	25–41	41–57	57–73	73–89	89–105	105–121	121–137	137–153	153–169
Frequency	42	48	47	40	41	21	5	2	2	2

(b) Also compute the quartiles and the quartile-deviation from the above data. *(M. Sc. Ag. Eco. Agra, 1965)*

Q. 9. The following table gives the yield of wheat from 10 equal plots :

Plot No.	1	2	3	4	5	6	7	8	9	10	Total
Yield in kgms.	60	40	50	45	60	55	65	50	65	55	45

If the area of each plot is 242 square yards, find the average yield per acre ? Also calculate the coefficient of variation (1 acre =4840 sq. yds.). *(U. P. Board 1963)*

Q. 10. The following table gives the marks obtained by some students. Calculate the arithmatic-mean and the mode :

Marks	0–10	10–20	20–30	30–40	40–50
Frequency	3	13	18	12	5

(U. P. Board, 1963)

Q. 11. An analysis of the monthly wages paid to the workers in two firms A and B belonging to the same industry, gives the following results :

	Firm A	Firm B
No. of wage earners :	586	648
Average monthly wages :	Rs. 52·5	Rs. 47·5
Var. of the distr. of wage ;	100	121

(a) Which firm pays out the larger amount as monthly wages to the workers ?

(b) In which firm A or B, is there greater variability in individual wages ?

(c) What is the measure of average monthly wages of all the workers in the two firms A and B taken together ? *(I. A. S. 1951)*

Q. 12. Following is the distribution of marks secured by 139 candidates. Calculate the S. D. and the coefficient of variation ?

Marks (x)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	Total
Frequency (f)	5	10	20	40	30	20	10	4	139

Q. 13. Write short notes on the following :

(i) Classification,

(ii) Histogram,

(iii) Ogive,

(iv) Frequency polygon and frequency curve,

(v) Median and Mode.

Q. 14. The scores of two golfers for 10 rounds each are :

A : 9, 17, 14, 13, 15, 10, 11, 13, 13, 15
B : 8, 15, 11, 11, 9, 12, 11, 10, 9, 14

Which player may be regarded as more consistent ?

Q. 15. The following table gives the heights and dry weights of 17 plants. Calculate the mean and S. D. for both the characters and find which of these two varies more ?

Heights in cms. (x) :
50, 44, 54, 46, 64, 37, 17, 62, 45, 60, 44, 57, 63, 68, 45, 71, 50

Weights in gms. (y) :
44, 25, 44, 26, 41, 21, 60, 52, 15, 55, 42, 58, 37, 36, 40, 58, 36

Q. 16. In any two series, where d_1 and d_2 represent the deviations from the assumed means $A_1=50$, $A_2=154$ have $n_1=10$, $n_2=9$, $\Sigma d_1=5{\cdot}8$, $\Sigma d_1^2=755{\cdot}66$, $\Sigma d_2=3$, $\Sigma d_2^2=55117{\cdot}0$. Calculate the coefficients of variation for the two series ?

Answers for Exercise 1.

Q. 2. (b)

Class	*Freq.*
5–10	7
10–15	6
15–20	5
20–25	9
25–30	4
30–35	4
35–40	3
40–45	2

The boundary line cases are kept in the classes where they are as lower limits.

Q. 7. Statistics ; (median is higher).

Q. 8. (a) Md.=53·08.

(b) Q_1=31·916, Q_3=77·390, Q. D.=22·737

Q. 9. 1090 Kgms./acre, C. V.=14·47.

Q. 10. A. M.=25·59, Mo.=24·545.

Q. 11. (a) B, (b) B, (c) 49·9 Rs.

Q. 12. S. D.=15·6, C. V.=39·2.

Q. 14. B.

Q. 15. $\bar{x}$=55·65 cms., $\bar{y}$=40·59 gms., S. D. (x)=11·71 cms., S. D. (y)=13·45 gms. weights vary more.

Q. 16. C. $V._1$=17·1, C. $V._2$=50·7.

Chapter IV

Correlation and Regression

Meaning of correlation : In all the previous chapters, we have dealt with elementary statistical analysis involving only a single measurement (variate X). Many statistical problems arise, in which the sample consists of pairs of measurements. For example, consider the behaviour of two variables, volume and pressure of a gas at a fixed temperature. In this case, we observe the relation that the pressure decreases as the volume increases, and if the pressure increases the volume decreases.

Another example may be taken as the behaviour of volume of a gas at a fixed pressure. In this case, the volume increases with the increase in the temperature and with the decrease of temperature, the volume also decreases.

In the former example, the changes in the two variables are in the opposite directions and in the later case, the changes are in the same direction. There is a number of phenomena of this type in exact and social sciences, where it is observed that the change in one variable is on an average accompanied by a change in the another variable, either in the same or in the opposite direction. The variables, which show such type of relations are said to be correlated, and the phenomenon itself is called *"Correlation"*.

Definition : The two variables are said to be correlated when the change in one variable is on an average accompanied by the change in the other variable, in the same or opposite directions.

Direction of correlation : The correlation is said to be positive or negative accordingly as the changes in the two variables are in the same or reverse directions.

For example, the heights and weights of a number of persons show the +ve correlation. Another examples of +ve correlations are—

the radius and circumference of a circle, dividend declared by a company and the market price of its share, the total cultivable area and the area under wheat etc. On the other hand, the price and demand of a certain commodity, area under fodder and non-fodder crops of a locality, the total income and the proportion of it spent on food, are the examples of −ve correlations.

Degree of correlation : There are three degrees of correlation, namely *perfect*, *limited* and *absence*. The correlation will be *perfect* when the ratio of the changes in the two variables is the same. The radius and circumference of a circle show the +ve perfect correlation. There are some instances where the correlation does not exist as in the case of height of a growing tree and that of a house, since the change in one is not accompanied by the change in the other. The degree of correlation may be limited. In this case, the changes in the two variables are not in a definite ratio or the percentage of change in one may not be equal to the percentage change in the other. For example, a 10% decrease it price may result in a 15% rise in demand for a certain commodity.

Thus, when the percentage changes in the two variables are the same, the correlation is perfect and when the percentage changes are not equal, the correlation is limited. According to the direction and degree of correlation, it may be—

(1) perfect positive,
(2) limited degree of positive,
(3) no correlation,
(3) llmited degree of negative and
(5) perfect negative.

1. Perfect positive correlation : The correlation will be perfectly +ve if an increase (or decrease) in one variable is accompanied by an increase (or decrease) in a perfectly definite ratio in the other variable. The radius and the circumference of a circle, the volume and the temperature of a gas at a fixed pressure, are the examples of perfect +ve correlation.

2. Limited degree of +ve correlation : If the increase (or decrease) in one variable is on an average accompanied by an increase (or decrease) in the other variable, but not necessarily in the same ratio, it is called limited degree of +*ve* correlation. Since, as the total cultivable area increases, the area under wheat also increases but not necessarily in the same ratio.

3. No correlation : The two variables are said to be uncorrelated when the change in one does not effect the other. As an example of this type is the number of radio sets produced and the total number of births recorded during a certain period.

4 Limited degree of −ve correlation : If the increase (or decrease) in one variable is on an average accompanied by a decrease (or inbrease) in the other variable, but not necessarily in the some ratio, it is called the limited degree of negative correlation. The area under the fodder and non-fodder crops will be an example of the limited degree of −*ve* correlation. Since, as the area under fodder crops increases, that of non-fodder crops fall but not necessarily in the same ratio.

5. Perfect −ve correlation : The correlation will be perfectly −ve, if an increase (or decrease) in one variable is accompanied by a decrease (or increase) in the other variable in a definite ratio. The two variables, volume and pressure of a gas at a fixed temperature show the perfect −ve correlation.

Measures of correlation :

(*1*) *The scatter diagram,*
(*2*) *Pearsons's coefficient of correlatio,n*
(*3*) *The regression line.*

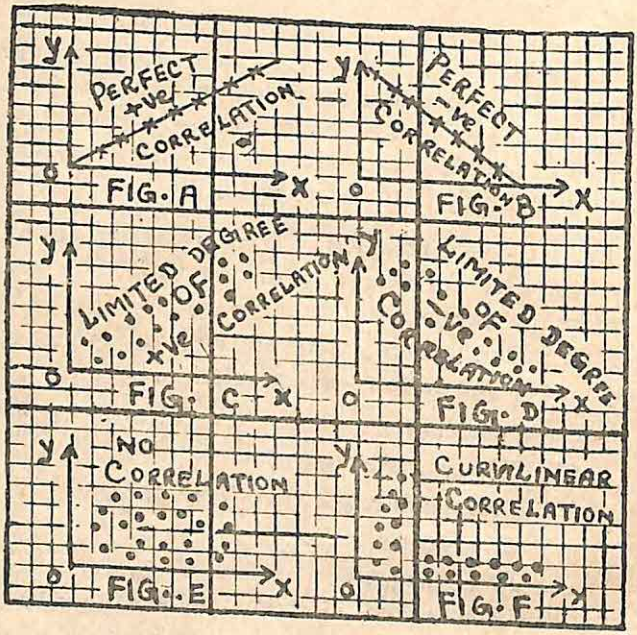
1. The scatter diagram : It is a graphic method of determining the degree and direction of the correlation. The method lies in plotting the (x, y) points on the graph paper. Merely an eye-inspection of the scatter-diagram (the diagram obtained by these points) is sufficient to decide the degree and direction of the correlation. If—

(i) all the points on the scatter diagram lie on a straight line with +ve slope (starting from the lower left hand corner to the upper right hand corner), the correlation will be perfect +ve (as shown in fig. A),

(ii) all the points on a scatter diagram lie along a straight line with −ve slope (*i. e.* the line starting from the upper left hand corner to the lower right hand corner) the correlation is perfect −ve (as shown in fig. B.),

(iii) the points on a scatter diagram seem to cluster about the diagonal with +ve slope, the correlation will be said as limited degree of +ve correlation (as shown in fig. C.),

(iv) the points on a scatter diagram seem to cluster about the diagonal with −ve slope, the correlation will be said as limited degree of −ve correlation (as shown in fig. D),



(v) the points on a scatter diagram take roughly the elliptical or the circular shape, then the two variables are said uncorrelated (as shown in the fig. E),

(vi) the scatter diagram takes the shape of a curve, the correlation will be curvilinear (as shown in fig. F), while in the cases (i)–(v) the correlation is linear and here we are concerned with linear correlation only.

Exp. 1. Point out the shortcomings of scatter diagram in measuring the correlation ? Suggest a measure free from these shortcomings ? Illustrate, how it is computed ?

Sol. The scatter diagram gives a clear idea about the direction of the correlation but as regards the degree of correlation, it can only tell us, whether the correlation is perfect or limited. In the case of limited degree of correlation, we required to know the extent of correlation to which a scatter diagram fails to answer. Moreover, the study through diagrams depends upon the accuracy and skill of the observer which is an uncertain factor. Hence our study and the conclusion drawn from it, may not be a reliable one. Due to this reason, we are in need of some arithmatical description which can provide a pure number free from the units in which the variables have been expressed and may indicate the direction of correlation also. The value of such measure should be -1 for perfect $-$ve correlation, $+1$ for perfect $+$ve correlation, and zero for non-correlation. The $-$ve values lying between 0 and -1 will indicate the limited degree of $-$ve correlation and $+$ve values lying between 0 and $+1$ will indicate the limited degree of $+$ve correlation. One such a measure is the coefficient of correlation, which is also called Pearson's coefficient of correlation or the product moment correlation coefficient. This Pearson's coefficient of correlation is generally denoted by '*r*' and is given bythe relation

$$r=\frac{\text{cov}(X, Y)}{\sigma_X\cdot\sigma_Y},\ \text{where cov.}(X, Y)=\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n}$$

i. e. the mean of the products of the deviations of the variates X and Y from their respective means $\overline{X}$ and $\overline{Y}$.

$$\sigma_X=\text{S. D. of } X=\sqrt{\left(\frac{\Sigma(X-\overline{X})^2}{n}\right)}$$

$$\text{and}\quad \sigma_Y=\text{S. D. of } Y=\sqrt{\left(\frac{\Sigma(Y-\overline{Y})^2}{n}\right)}$$

$$\text{Hence,}\quad r=\frac{\Sigma(X-\overline{X})(Y-\overline{Y})/n}{\sqrt{\left(\frac{\Sigma(X-\overline{X})^2}{n}\right)}\times\sqrt{\left(\frac{\Sigma(Y-\overline{Y})^2}{n}\right)}}$$

$$\therefore\quad r=\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\{\Sigma(X-\overline{X})^2\}}\cdot\sqrt{\{\Sigma(Y-\overline{Y})^2\}}}$$

If the deviations are taken from the assumed means, the formula takes the form—

$$r=\frac{\Sigma\,\xi\eta-\Sigma\,\xi\;\Sigma\eta/n}{\sqrt{\left\{\left(\Sigma\,\xi^2-\frac{(\Sigma\,\xi)^2}{n}\right)\right\}}\sqrt{\left\{\left(\Sigma\,\eta^2-\frac{(\Sigma\,\eta)^2}{n}\right)\right\}}}$$

Where ξ, η are the deviations of the variates x and y from their respective assumed means A_x and A_y.

This formula is of great importance when the deviations from the actual means are fractional. In the computation of correlation coefficient, the main steps are given below:—

(a) (i) choose an assumed mean A_x for the variate X,
 (ii) obtain the deviations $\xi=X-A_x$,
 (iii) calculate ξ^2,

(b) (i) choose the assumed mean A_y for the variate Y,
 (ii) obtain the deviations $\eta=Y-A_y$
 (iii) calculate the squares of η

(c) (i) compute the praduct $\xi\eta$,

For the illustration of this method, we compute r from the following examples—

Exp. (2) : The percent carotene content of wheat (X) and the percent carotene content of flour (Y) for 10 varieties of wheat were found to be as follows—

X :	1.18	2.13	1.41	1.42	1.50	1.25	1.65	1.24	1.48	1.35
Y :	2.39	3.11	2.15	1.96	2.02	1.76	2.10	2.12	2.28	1.86

Find the correlation coefficient for X and Y ?

Sol. : To compute the correllation coefficient for X and Y, we prepare the following table—

$A_x = 1.50$ $A_y = 2.02$

Variety No.	X	$\xi = X - A_x$	ξ^2	Y	$\eta = Y - A_y$	η^2	$\xi\eta$
1.	1.18	−.32	0.1024	2.39	+.37	0.1369	−0.1184
2.	2.13	+.63	0.3969	3.11	+1.09	1.1881	+0.6867
3.	1.41	−.09	0.0081	2.15	+.13	0.0169	−0.0117
4.	1.42	−.08	0.0064	1.96	−.06	0.0036	+0.0048
5.	1.50	0	0	2.02	0	0	0
6.	1.52	−.25	0.0625	1.76	−.26	0.0676	+0.0650
7.	1.65	+.15	0.0225	2.10	+.08	0.0064	+0.0120
8.	1.24	−.26	0.0676	2.12	+.10	0.0100	−0.0260
9.	1.48	−.02	0.0004	2.28	+.26	0.0676	−0.0052
10	1.35	−.15	0.0225	1.86	−.16	0.0256	+0.0240
Totals	$n=10$	−.39 $=\Sigma\xi$	0.6893 $=\Sigma\xi^2$	$n=10$	+1.55 $=\Sigma\eta$	1.5227 $=\Sigma\eta^2$	0.6312 $=\Sigma\xi\eta$

Now the correlation coefficient is given by the formula—

$$r = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi\Sigma\eta}{n}}{\sqrt{\left[\Sigma\xi^2 - \dfrac{(\Sigma\xi)^2}{n}\right]}\sqrt{\left[\Sigma\eta^2 - \dfrac{(\Sigma\eta)^2}{n}\right]}}$$

$$= \frac{0.6312 - \dfrac{-.39 \times 1{\cdot}55}{10}}{\sqrt{\left[{\cdot}6893 - \dfrac{(-{\cdot}39)^2}{10}\right]}\sqrt{\left[1.5227 - \dfrac{(1.55)^2}{10}\right]}}$$

$$= \frac{0.63120 + .06045}{\sqrt{(0.67409)}\sqrt{(1.28245)}}$$

$$=\frac{0.69165}{\sqrt{(0.8644867205)}}=\frac{0.69165}{0.92977}=+0.7348$$

$$\therefore \quad r=+0.7348.$$

(2) Pearson's coefficient of correlation :

Exp. (3) : In determining volume of red cells in defibrinated beef blood by a certain method, the following results were obtained—

Percent of whole blood in mixture (X)	100	100	100	90	90	90	80	80	80	50	50	50
Percent of volume occupied by red cells (Y)	40·5	44·5	46·0	35·5	40·5	41·0	32·0	36·0	36·5	20·0	22·0	23·0

Find the product moment coefficient for X and Y ?

Sol. : The computation of product moment coefficient is made in the following tabular form—

$A_x = 80$ $A_y = 40 \cdot 5$

S. No.	X	$\xi = X - A_x$	ξ^2	Y	$\eta = Y - A_y$	η^2	$\xi\eta$
1.	100	+20	400	40·5	0	0	0
2.	100	+20	400	44·5	+4·0	16·00	+80·00
3.	100	+20	400	46·0	+5·5	30·25	+110·00
4.	90	+10	100	35·5	−5·0	25·00	−50·00
5.	90	+10	100	40·5	0	0	0
6.	90	+10	100	41·0	+0·5	00·25	+5·00
7.	80	0	0	32·0	−8·5	72·25	0
8.	80	0	0	36·0	−4·5	20·25	0
9.	80	0	0	36·5	−4·0	16·00	0
10.	50	−30	900	20·0	−20·5	420·25	+615·00
11.	50	−30	900	22·0	−18·5	342·24	+555·00
12.	50	−30	900	23·0	−17·5	306·25	+525·00
Totals	$n=12$	$0=\Sigma\xi$	$4200 = \Sigma\xi^2$	$n=12$	$-68 \cdot 5 = \Sigma\eta$	$1248 \cdot 74 = \Sigma\eta^2$	$+1840 \cdot 00 = \Sigma\xi\eta$

Hence we have

$$r = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi . \Sigma\eta}{n}}{\left[\Sigma\xi^2 - \dfrac{(\Sigma\xi)^2}{n}\right]\sqrt{\left[\Sigma\eta^2 - \dfrac{(\Sigma\eta)^2}{n}\right]}}$$

$$= \frac{1840 - \dfrac{0 \times (-68.5)}{12}}{\sqrt{\left[4200 - \dfrac{(0)^2}{12}\right]}\sqrt{\left[1248.74 - \dfrac{(-68.5)^2}{12}\right]}}$$

$$= \frac{1840}{\sqrt{(4200)}\sqrt{(857.7192)}} = \frac{1840}{\sqrt{(3602420.64)}} = \frac{1840}{1898.04}$$

$$= +0.9694$$

$\therefore r = +0.97.$

Ex. (4) A research worker observed the following fresh weights and dry weights in a sample of his experimental material—

Observation No.	1	2	3	4	5	6	7	8	9	10
Fresh wt. (oz)	8	6	10	5	12	2	20	15	14	18
Dry wt. (oz)	3	2	2	2	4	1	5	4	3	4

(a) Calculate the correlation between the two characters and interpret them ?

(b) Represent the data in a scatter diagram ?

(M. Sc. Ag. Agra 1963)

Sol. :

(a) Taking fresh weights by X and dry weights by Y, we prepare the table—

$A_x = 11$ $A_y = 3$

Obs. No.	X	$\xi = X - A_x$	ξ^2	Y	$\eta = Y - A_y$	η^2	$\xi\eta$
1.	8	-3	9	3	0	0	0
2.	6	-5	25	2	-1	1	$+5$
3.	10	-1	1	2	-1	1	$+1$
4.	5	-6	36	2	-1	1	$+6$
5.	12	$+1$	1	4	$+1$	1	$+1$
6.	2	-9	81	1	-2	4	$+18$
7.	20	$+9$	81	5	$+2$	4	$+18$
8.	15	$+4$	16	4	$+1$	1	$+4$
9.	14	$+3$	9	3	0	0	0
10.	18	$+7$	49	4	$+1$	1	$+7$
Totals	$n = 10$	$0 = \Sigma\xi$	$308 = \Sigma\xi^2$	$n = 10$	$0 = \Sigma\eta$	$14 = \Sigma\eta^2$	$60 = \Sigma\xi\eta$

Now we have—

$$r=\frac{\Sigma\xi\eta-\frac{\Sigma\xi.\Sigma\eta}{n}}{\sqrt{\left[\Sigma\xi^2-\frac{(\Sigma\xi)^2}{n}\right]}\sqrt{\left[\Sigma\eta^2-\frac{(\Sigma\eta)^2}{n}\right]}}$$

$$=\frac{60-\frac{0\times0}{10}}{\sqrt{\left[308-\frac{(0)^2}{10}\right]}\sqrt{\left[14-\frac{(0)^2}{10}\right]}}$$

$$=\frac{60}{\sqrt{(308)}\sqrt{(14)}}=\frac{60}{14\times\sqrt{22}}=\frac{30}{7\times4.69}=\frac{30}{32.83}=0.9137$$

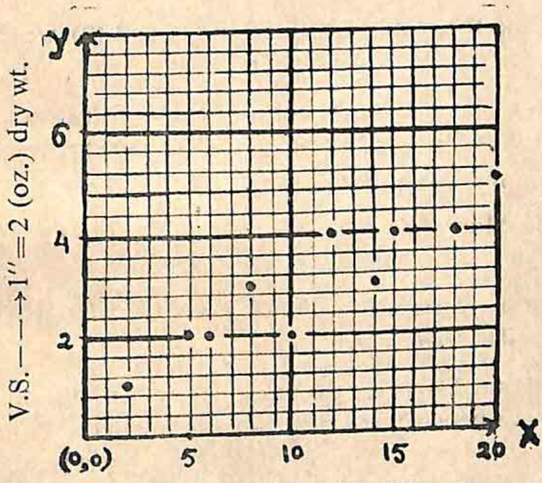
$\therefore\ r=+0.9137.$

Interpretation : The two characters show a high degree of correlation. It means that there is nearly a linear relationship between the fresh weights and the dry weights of the sample.

(b) The scatter diagram for the fresh and dry weights :

In the scatter diagram, the points (X, Y) cluster around the diagonal with +ve slope, which shows a high degree of +ve correlation in the fresh and dry weights.

Scatter Diagram



H.S.→1 small div.=1(oz.) Fresh wt.

Exp. (5) : Show that the coefficient af correlation lies between -1 and $+1$. *(M. Sc. Agra, 1948, M. Sc. Ag. Agra 1965)*

Sol. : Consider the quantity

$$S=\frac{1}{2n}\,\Sigma\left[\frac{X-\bar{X})}{\sigma_X}-\frac{(Y-\bar{Y})}{\sigma_Y\,\sigma_X}\right]^2 \geqslant 0$$ Since the sum of squares is always +ve.

Where $\bar{X}$, $\bar{Y}$ **are the means** of the variates X and Y respectively and σ_X, σ_Y their S. Ds.

or $$\frac{1}{2n}\,\Sigma\left[\frac{(X-x)^2}{\sigma_X{}^2}+\frac{(Y-\bar{Y})^2}{\sigma_Y{}^2}-2\,\frac{(X-\bar{X})(Y-\bar{Y})}{\sigma_X\,\sigma_Y}\right]\geqslant 0$$

$$S=\frac{1}{2}\left[\frac{\Sigma(X-\bar{X})^2}{n.\sigma_X{}^2}+\frac{\Sigma(Y-\bar{Y})^2}{n.{}_Y{}^2}-2.\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n.\sigma_X\sigma_Y}\right]\geqslant 0$$

or $\frac{1}{2}[1+1-2r]\geqslant 0$

or $1-r\geqslant 0$

$\therefore\ r\leqslant +1.$ ………(i)

Further, taking the quantity

$$S'=\frac{1}{2n}\,\Sigma\left[\frac{(X-\bar{X})}{\sigma_X}+\frac{(Y-\bar{Y})}{\sigma_Y}\right]^2\geqslant 0$$

which gives us

$$1+r\geqslant 0$$

$\therefore$ $r\geqslant -1.$ ………(ii)

Combining the results obtained in (1) and (2) we get,

$-1\leqslant r\leqslant +1.$ **Hence proved.**

Exp. (6) : Define rank correlation ? Illustrate, how will you complete it with the help of the following example—

Marks in mathematics (X) :	80 75, 72, 72, 70, 68, 62, 50
Marks in Music (Y) :	10, 12, 17, 17, 17, 22, 20, 23

Sol.

Def. : The correlation between the grades (ranks) of the variables is called '*rank correlation*'.

It is computed in the circumstances when we want to know the relationship between the proficiencies of a group of candidates in the two different subjects A and B or we wish to investigate the degree of agreement between the two judges who have graded the same individuals regarding the same characteristic. The rank correlation is given by the following formula [due to spearman]

$r=1-\frac{6\Sigma d^2}{n(n^2-1)}$, where $d\rightarrow$represents the difference between the ranks of the same individual,

$n \rightarrow$ the number of observations.

Method of ranking : First go through the series carefully and note the item with highest magnitude. This will be assigned rank 1, the next lower rank 2, and so on, ..., the item with lowest magnitude will be assigned the highest rank *i.e.* equal to the number of observations.

Consider the series—

78, 36, 25, 98, 75, where the 4th item has the highest magnitude so its rank will be 1, next lower item is 78 so its rank will be 2, then comes 75 with rank 3, 36 with rank 4 and 25 the lowest value with rank 5 (the highest rank). Now we summarize the result as below

Measurement (X) :	78,	36,	25,	98,	75
Rank (r) :	2,	4,	5,	1,	3

In the case when two or more items of (X) are equel, the rank awarded to each of them will be the A. M. of the ranks that they would have received if they had differed by a small quantity. For this, consider the series—

81, 78, 73, 69, 73, 78, 62, 73

Let us first suppose that all the values are different, then the rank assigned to 1st item is 1, (being highest value), to second item is 2, to sixth item is 3, to third item is 4, to fifth item is 5, to eighth item is 6, to fourth item is 7 and to the seventh item is 8. Here 78 has appeared twice, so the rank assigned to 78 will be$=\frac{2+3}{3}$ *i.e.* 25, 37 also appeared thrice so its rank will be $=\frac{5+4+6}{3}=5$; Now we summarize the result as follows—

Measurement (X) :	81,	78,	73,	69,	73,	78,	62,	73
Actual rank (*r*) :	1,	2.5,	5,	7,	5,	2.5,	8,	5,

The computation of the rank correlation is made as follows—

X	Rank of X	Y	Rank of Y	Rank diff. '*d*'	d^2
80	1	10	8	7	49·00
75	2	12	7	5·00	25·00
72	3·5	17	5	1·50	2·25
72	3·5	17	5	1·5	2·25
70	5	17	5	0	0
68	6	22	2	−4	16·00
62	7	20	3	−4	16·00
50	8	23	1	−7	49·00
Totals	—	—	—	—	159·50 $=\Sigma d^2$

Coefficient of rank correlation$=1-\dfrac{6\Sigma d^2}{n(n^2-1)}$

$$=1-\frac{6\times 159.50}{8(64-1)}=1-\frac{957}{504}$$

$\therefore$ $r=-0.898$, the high−ve value shows that a student best in Maths is worst in music.

Exp. (7) : The ranks of 15 candidates in a beauty context ranked by two judges were as follows—

(The two numbers within brackets denoting the ranks of the same candidate)

(6, 8), (7, 3), (8, 1), (9, 11), (10, 15), (11, 9), (12, 5), (13, 14), (14, 12), (15, 13), (1, 10), (2, 7), (3, 2), (4, 6), (5, 4). Show that the two judges do not differ in their opinions ?

Sol. The computation of the rank correlation is made as follows—

$d=2, -4, -7, 2, 5, -2, -7, 1, -2, -2, 9, 5, -1, 2, -1$

$d^2=4, 16, 49, 4, 25, 4, 49, 1, 4, 4, 81, 25, 1, 4, 1$

$\Sigma d^2=272,\quad n=15$

$$\therefore\ r=\frac{6\ \Sigma d^2}{n(n^2-1)}$$

$$=1-\frac{6\times 272}{15\,(225-1)}$$

$$=1-\frac{1632}{3360}$$

$$\therefore\ r=+0.5142$$

The +ve value of the rank correlation coefficient shows that the judges do not differ in their opinions.

(3) Regression : The correlation theory deals with the relationship exhibited by two variables *i. e.* whether the change in one variable tends to bring the change in the other in the same or inverse direction. When such relation is exhibited by the two variables, then it is also intended to calculate the amount of change that will on an average take place in the other variable when a unit change has occured in one. This average change is given by the *regression-coefficient*.

Using the regression coefficient, we can give the expected value (average value) of a variate corresponding to a fixed value of the other variate. The variate which is kept fixed is called the *independent variable* and the other (which is to be estimated) is called *dependent variable*. Of the two variables, we choose that variable as independent variable whose effect produces the correlation. For example, consider the rainfall and yield of rice which are often correlated. The correlation between the two is the result of the influence of rainfall on yield of rice and not the result of the influence of yield of rice on rainfall. Hence the yield of rice will be taken as the dependent variable and the rainfall as the independent variable. Sometimes both the variables may take both of the roles. It happens when the correlation between them is the result of the influence of a third factor. For example, the +ve correlation between the yield of rice/acre and of jute/acre is due to the fact that the two are related to the amount of rainfall. Thus the regression concept is concerned with the way in which the changes in one variable depend upon the simultaneous changes in the other variable (called the independent variable). In general, if Y is the dependent and X the independent factor, the recorded value of Y for any value of X will show the ordinary variation and thus it will cover a range-readings. *The regression function* is that which expresses the average value (that might be expected) of one variable for any given value of the independent factor.

We are concerned with the linear regression only *i. e.* when the regression function represents a straight line called the regression line. The regression line of Y on X is defined by the equation :

$$Y-\bar{Y}=r\,\frac{\sigma_Y}{\sigma_X}\,(X-\bar{X}) \qquad \ldots\ldots\text{(i)}$$

where $\bar{X}$ and $\bar{Y}$ are the means of X and Y variables respectively and σ_X, σ_Y their standard deviations ; r is the coefficient of correlation between the two variables.

Similarly, the regression line of X on Y may be

$$X-\bar{X}=r\,\frac{\sigma_X}{\sigma_Y}\,(Y-\bar{Y}) \qquad \ldots\ldots\text{(ii)}$$

In equations (i) and (ii), the variables on the right hand side are the independent variables and those on the L. H. S. are the dependent. The relation (i) is used to estimate Y corresponding to a given value of X and (ii) is used to estimate the value of X for a given value of Y. We can replace $r\,\frac{\sigma_Y}{\sigma_X}$ by b_{YX} in (i) and $r\,\frac{\sigma_X}{\sigma_Y}$ by b_{XY} in (ii), which are called the regression-coefficients of Y on X and of X on Y respectively. The regression-coefficient b_{YX} is computed as follows :

$$b_{YX}=r\,\frac{\sigma_Y}{\sigma_X}$$

$$=\frac{\Sigma\xi\eta-\frac{\Sigma\xi\Sigma\eta}{n}}{\sqrt{\left[\Sigma\xi^2-\frac{(\Sigma\xi)^2}{n}\right]}\sqrt{\left[\Sigma\eta^2-\frac{(\Sigma\eta)^2}{n}\right]}}\times\frac{\sqrt{\left[\Sigma\eta^2-\frac{(\Sigma\eta)^2}{n}\right]}}{\sqrt{\left[\Sigma\xi^2-\frac{(\Sigma\xi)^2}{n}\right]}}$$

$$\therefore\ b_{YX}=\frac{\Sigma\xi\eta-\frac{\Sigma\xi\Sigma\eta}{n}}{\left[\Sigma\xi^2-\frac{(\Sigma\xi)^2}{n}\right]},$$

where ξ, η are the deviations of X and Y from their assumed means, $n \longrightarrow$ is the no. of observations.

Similarly,

$$b_{XY}=\frac{\Sigma\xi\eta-\frac{\Sigma\xi\Sigma\eta}{n}}{\left[\Sigma\eta^2-\frac{(\Sigma\eta)^2}{n}\right]}$$

Exp. 8. (a) Give the relation between the regression coefficients and the correlation coefficient ? Also give the difference between them ?

(b) Point out the difference between correlation and regression theory ?

(c) How the regression lines give an idea of the correlation ?

(d) What is the point of intersection of the 2 regression lines ?

Sol. (a) The correlation coefficient is the G. M. of the regression coefficients. As we know that

$$b_{YX} = r\,\frac{\sigma_Y}{\sigma_X}$$

$$\text{and}\quad b_{XY} = r\,\frac{\sigma_X}{\sigma_Y}.$$

$$\text{Hence, } b_{YX}\,.\,b_{XY} = r^2.\frac{\sigma_Y}{\sigma_X}.\frac{\sigma_X}{\sigma_Y} = r^2$$

$$\therefore\quad r = \sqrt{(\,b_{YX}\,.\,b_{XY}\,)}.$$

The correlation coefficient is a measure of direction and extent of correlation between the two variables. It indicates whether the change in one variable tends to bring a change in the other variable in the same or reverse direction. While the regression coefficient gives the average change in one variable corresponding to a unit change in the other variable. Thus b_{YX} gives an average change in Y for a unit change in X.

(b) The fundamental difference between the correlation and regression theory is that both the variables are considered as random in the former case while in the later, one variable is independent and the other is dependent.

(c) When the two regression lines coincide, the correlation is perfect and when they cut at rightangle, the correlation is zero. If the two lines diverge and intersect each other, the correlation is limited. As the divergence between the two regression lines increases, the correlation decreases.

(d) The two regression lines intersect at the point $(\bar{X}, \bar{Y})$ i. e. they pass through the means.

Note. $b_{YX} = \tan\theta$, where θ is the angle which the regression lines of Y on X makes with the X-axis.

and $b_{XY} = \tan\phi$, where ϕ is the angle which the regression line of X on Y makes with the Y-axis.

Exp. No. 9. (a) Find the two regression lines from the following data :—

(b) Plot the lines on graph and give an idea of 'r' ?

Fresh wts. (gms.) (x) : 25, 27, 26, 29, 34, 35,
Dry wts. (gms.) (y) : 2, 5, 7, 9, 19, 17,

Sol. : (a) For the computation of the two regression lines, we prepare the following table :

$A_x=30$ $A_y=10$

X	$\xi=X-A_x$	ξ^2	Y	$\eta=Y-A_y$	η^2	$\xi\eta$
25	−5	25	2	−8	64	40
27	−3	9	5	−5	25	15
26	−4	16	7	−3	9	12
29	−1	1	9	−1	1	1
34	+4	16	19	+9	81	36
35	+5	25	17	+7	49	35
Totals	$-4=\Sigma\xi$	$92=\Sigma\xi^2$	$n=6$	$-1=\Sigma\eta$	$229=\Sigma\eta^2$	$139=\Sigma\xi\eta$

$$\bar{X}=A_x+\frac{\Sigma\xi}{n}=30+\frac{-4}{6}=30-0\cdot67=29\cdot33$$

$$\bar{Y}=A_y+\frac{\Sigma\eta}{n}=10+\frac{-1}{6}=10-0\cdot17=9\cdot83$$

$$b_{XY}=\frac{\Sigma\xi\eta-\frac{\Sigma\xi\Sigma\eta}{n}}{\left[\Sigma\xi^2-\frac{(\Sigma\xi)^2}{n}\right]}=\frac{139-\frac{(-4)X(-1)}{6}}{\left[92-\frac{(-4)^2}{6}\right]}$$

$$=\frac{139-0\cdot67}{92-2\cdot67}=\frac{138\cdot33}{89\cdot33}=1\cdot55 \text{ approx.}$$

and $$b_{XY}=\frac{\Sigma\xi\eta-\frac{\Sigma\xi\Sigma\eta}{n}}{\left[\Sigma\eta^2-\frac{(\Sigma\eta)^2}{n}\right]}=\frac{139-\frac{(-4)\,(-1)}{6}}{\left[229-\frac{(-1)^2}{6}\right]}$$

$$=\frac{139-0\cdot67}{229-0\cdot17}=\frac{138\cdot33}{228\cdot83}=0\cdot60 \text{ approx.}$$

Therefore, the regression line of Y on X, will be

$$Y-\overline{Y}=b_{YX}\,(X-\overline{X})$$

or $$Y-9{\cdot}83=1{\cdot}55\,(X-29{\cdot}33) \quad \ldots \quad \ldots \quad \ldots(i)$$

and that of X on Y will be

$$X-\overline{X}=b_{XY}\,(Y-\overline{Y})$$

or $$X-29{\cdot}33=0{\cdot}60\,(Y-9{\cdot}83) \quad \ldots \quad \ldots \quad \ldots(ii)$$

Hence the regression lines are $Y=1{\cdot}55X-35{\cdot}63$
and $X=0{\cdot}60Y+23{\cdot}43$

(b) To plot the graphs of the regression lines

$$Y-9{\cdot}83=1{\cdot}55\,(X-29{\cdot}33 \quad \ldots \quad \ldots \quad \ldots(i)$$

and $$X-29{\cdot}33=0{\cdot}60\,(Y-9{\cdot}83) \quad \ldots \quad \ldots \quad \ldots(ii)$$

we need at least two points for each. One of the points will be

$$M\equiv(\overline{X},\,\overline{Y})\equiv(29{\cdot}33,\,9{\cdot}83)$$

as the lines always pass through the point $(\overline{X}, \overline{Y})$. For another point of equation (i).

Put $X=0$ in (i), to get

$$Y=35{\cdot}63.$$

Hence the point say $A\equiv(0.\ 35{\cdot}63)$.

If we join M and A, we get the graph of equation (i).

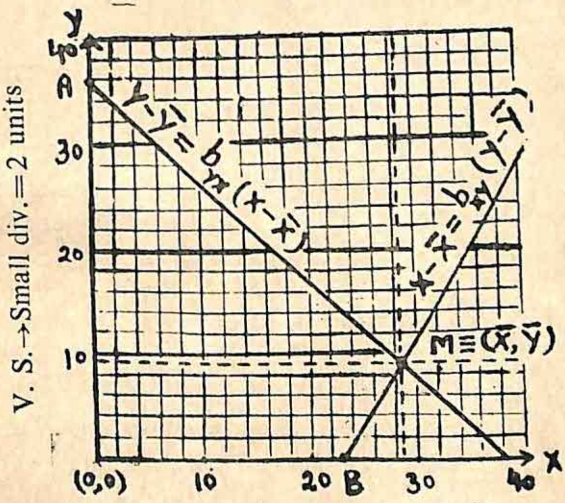
Similarly, put $Y=0$ in (ii), to get

$$X=23{\cdot}43.$$

Hence the point say $B\equiv(23{\cdot}43,\ 0)$.

If we join M and B, we get the graph of equation (ii).

Graphs for Regression Lines



H. S.→1 Small div.=2 units

The graphs of the two regression lines show the limited degree of positive correlation between X and Y.

The significance of correlation coefficient: In the case of *large samples*, (i. e. $n > 100$), to test the significance of the correlation coefficient, we compute the statistic

$$z=\frac{r}{1-r^2}\times\sqrt{(n-1)}, \text{ since S. E. } (r)=\frac{1-r^2}{\sqrt{(n-1)}}$$

which is a S. N. V.

Hence the properties of S. N. V. can be used to determine whether a correlation coefficient is significantly different from zero or not.

For illustration, consider the following example—

Exp. 10. A coefficient of correlation of 0·22 is obtained from a sample of 401 observations. Is this correlation significant ?

Sol. $z=\frac{0{\cdot}22}{1-0{\cdot}0484}\times 20$

$=\frac{4{\cdot}4}{0{\cdot}9516}$.

$\therefore$ $z=4{\cdot}6$ approx., which is greater than 1·96 the value of S. N. V. at 5 percent.

Hence the correlation coefficient is significantly different from zero at 5 percent level of significance.

In the case of *small samples* (i, e. $n < 100$), the statistic

$$\frac{r}{1-r^2}.\sqrt{(n-1)}$$

does not conform to the normal or other standard form. Therefore, we find a simple function of 'r' which conforms to some standard form. For this, the function of 'r' is

$$\frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)},$$

which is found to obey the 't' distribution with $(n-2)$ d. f.

For $\rho=0$ (population correlation coefficient), whatever be the size of the sample, we always calculate the statistic

$$t=\frac{r}{\sqrt{(1-r^2)}}.\sqrt{(n-2)},$$

and use the properties of 't' distribution.

The significance of the regression coefficient : To test the significance of b_{YX}, we compute the statistic

$t=\dfrac{b_{YX}-\beta}{\text{S. E. of } b_{YX}}$, where β is the corresponding population coefficient of regression,

and $\text{S. E. of } (b_{YX})=\sqrt{\left(\dfrac{\text{S. S.}_{\cdot Y}-b^2{}_{YX}\,.\,\text{S. S.}_{\cdot X}}{(n-2).\text{S. S.}_{\cdot X}}\right)}$

or $=\dfrac{\sigma_Y{}'}{\sqrt{(n-2)}\,.\,\sigma_X}$,

where $\text{S. S.}_{\cdot X}=\Sigma\xi^2-\dfrac{(\Sigma\xi)^2}{n}$, ξ is the deviation of X from assumed mean A_X

or $\text{S. S.}_{\cdot X}=\Sigma x^2$, where $x=(X-\overline{X})$

i. e. deviation from actual mean $\overline{X}$.

Similarly, $\text{S. S.}_{\cdot Y}=\Sigma\eta^2-\dfrac{(\Sigma\eta)^2}{n}$, η is the deviation of Y from assumed mean A_Y

or $\text{S. S.}_{\cdot Y}=\Sigma y^2$, where $y=(Y-\overline{Y})$

i. e. deviation from the actual mean $\overline{Y}$,

where $\sigma_Y{}'=\sqrt{(\sigma_Y{}^2-b^2{}_{YX}\cdot\sigma_X{}^2)}$.

This statistic conforms to '*t*' distribution with $(n-2)$ d. f.

Similarly, to test the the significance of b_{XY}, we compute the statistic

$t=\dfrac{b_{XY}-\beta}{\text{S. E. of } b_{XY}}$

where $\text{S. E. of } b_{XY}=\sqrt{\left(\dfrac{\text{S. S.}_{\cdot X}-b^2{}_{XY}\,.\,\text{S. S.}_{\cdot Y}}{(n-2)\,\text{S. S.}_{\cdot Y}}\right)}$

or $=\dfrac{\sigma_X{}'}{\sqrt{(n-2)}\,.\,\sigma_Y}$

where $\sigma_X{}'=\sqrt{(\sigma_X{}^2-b^2{}_{XY}\cdot\sigma_Y{}^2)}$.

Exp. 11. The thickness of 20 annual rings of a tree and the corresponding annual rainfall were found to be correlated with a coefficient of $+0{\cdot}47$. Is this correlation significant ?

Sol. We have $r=0{\cdot}47$

$n=20.$

So we compute the statistic

$$\begin{aligned} t &= \frac{r}{\sqrt{(1-r^2)}} \cdot \sqrt{(n-2)} \\ &= \frac{0{\cdot}47}{\sqrt{[1-(0{\cdot}47)^2]}} \times \sqrt{18} \\ &= \frac{0{\cdot}47}{\sqrt{(0{\cdot}7791)}} \times \sqrt{18} \\ &= 0{\cdot}47 \times \sqrt{\left(\frac{18}{0{\cdot}7791}\right)} \\ &= 0{\cdot}47 \times \sqrt{(23{\cdot}103581)} \\ &= 0{\cdot}47 \times 4{\cdot}806 \\ &= 2{\cdot}25896. \end{aligned}$$

$$\therefore \quad t \simeq 2{\cdot}26, \quad t_{\cdot 05}(18)=2{\cdot}101.$$

Conclusion : This calculated value of 't' is greater than the tabulated value of t at 5 percent level of significance for 18 d. f. Hence the correlation coefficient is significant at 5 percent level.

Exp. (12) : The following data are for the amount of water applied in inches and yield of alfalfa in tons/acre—

Water (X) : 12, 18, 24, 30, 36, 42, 48

Yield (Y) : 5·27, 5·68, 6·25, 7·21, 8·20, 8·67, 8·42. Find the regression of yield on water ? Assuming that the relation between the two is linear, calculate the expected yield when the amount applied is 20 inches. *(M. Sc. Ag. Agra, 1959)*

Sol. : Here the water is independent variable and the yield is dependent variable and so they are represented by X and Y respectively. To compute the regression line, the following table is prepared—

X	$\xi = X - A_x$ $A = 30$	ξ^2	Y	$\eta = Y - A_y$ $A = 7{\cdot}08$	η^2	$\xi\eta$
12	−18	324	5·27	−1·81	3·2761	32·58
18	−12	144	5·68	−1·40	1·9600	16·80
24	−6	36	6·25	−0·83	0·6889	4·98
30	0	0	7·21	+0·13	0·1690	0
36	+6	36	8·20	+1·12	1·2544	6·72
42	+12	144	8·67	+1·59	2·5281	19·08
48	+18	324	8·42	+1·34	1·7956	24·12
Totals	$0 = \Sigma\xi$	$1008 = \Sigma\xi^2$	$n = 7$	$+0{\cdot}14 = \Sigma\eta$	11·6721	104·28

$$\bar{X} = A_x + \frac{\Sigma\xi}{n} = 30 + \frac{0}{7} = 30 \text{ inches}$$

$$\bar{Y} = A_y + \frac{\Sigma\eta}{n} = 7{\cdot}08 + \frac{0{\cdot}14}{7} = 7{\cdot}08 + 0{\cdot}02 = 7{\cdot}10 \text{ tons.}$$

The regression of yield on water is given by the equation—

$$Y - \bar{Y} = b_{YX}(X - \bar{X}), \text{ where } b_{YX} = \frac{\Sigma\xi\eta - \dfrac{\Sigma\xi\Sigma\eta}{\eta}}{\left[\Sigma\xi^2 - \dfrac{(\Sigma\xi)^2}{n}\right]} = \frac{104{\cdot}28 - 0}{1008 - 0}$$

$$= 0{\cdot}1035$$

$\therefore$ $$Y-7{\cdot}10=0{\cdot}1035(X-30)$$

or $$Y=0{\cdot}1035X+3{\cdot}995$$

For $X=20''$, the expected yield will be

$$Y_e=\overline{Y}+b_{YX}(20-30)$$

or $$Y_e=7{\cdot}10+0{\cdot}1035\times 10$$

or $$Y_e=7{\cdot}100+1{\cdot}035$$

$\therefore$ $$Y_e=8{\cdot}135 \text{ tons/acre.}$$

Thus the most likely yield of alfalfa per acre when water is 20″ is 8·135 tons.

Exp. (13) : Measurement of height of father (X inches) and of sons (Y inches) on thirty pairs of fathers and sons gave the following results to the nearest whole numbers,

$$\overline{X}=67\ \ \overline{Y}=68,\ S(X-\bar{x})^2=225,\ S(Y-\overline{Y})^2=256,$$
$$S(X-\bar{X})(Y-\bar{Y})=125$$

Calculate the regression of the height of son on the height of the father and predict the mean height of sons of fathers whose height is 70 inches ? *(M. Sc. Ag. Agra, 1956)*

Sol : We have $n=30$, $\Sigma x^2=225$, $\Sigma y^2=256$, $\Sigma xy=125$,
$$\bar{X}=67,\ \bar{Y}=68$$

The regression of the height of sons (Y) on the height of fathers (X) is given by

$$Y-\bar{Y}=b_{YX}(X-\bar{X}),\ \text{where } b_{YX}=\frac{S(X-\bar{X})(Y-\bar{Y})}{S(X-\bar{X})^2}=\frac{\Sigma xy}{\Sigma x^2}$$
$$=\frac{125}{225}=0{\cdot}56$$

or $$Y-68=0{\cdot}56(X-67)$$

or $$Y=0{\cdot}56X3+0{\cdot}48$$

To predict the mean height of sons of fathers whose height is 70 inches, we have

$$Y_e=\bar{Y}+b_{yx}(X-\bar{X})$$

or $$Y_e=68+0{\cdot}56(70-67)$$

or $Y_e=68+1{\cdot}68=69{\cdot}68$ inches. It is the expected mean height of sons of those fathers whose height is 70″. **Ans.**

Exp. (14) : From 15 plots of cotton, the number of plants per plot (X) and the corresponding yield (Y) were recorded and the following quantities obtained—

$$S(X-\bar{X})^2=175{\cdot}6, \quad S(Y-\bar{Y})^2=79{\cdot}6$$

$$S(X-\bar{X})(Y-\bar{Y})=102{\cdot}4, \quad \bar{Y}=11{\cdot}4, \; \bar{Y}=6{\cdot}4$$

Calculate the regression of yield on the no. of plants and test its significance. What will be the equation of the regression line ?

(M. Sc. Ag. Agra, 1960)

Sol. : Here we take the yield (Y) as the dependent variable and the no. of plants/plot (X) as the independent variable. To find the regression of yield on the no. of plants/plot, we compute

$$b_{YX}=\frac{(X-\bar{X})(Y-\bar{Y})}{S(X-\bar{X})^2}=\frac{\Sigma xy}{\Sigma x^2}=\frac{102{\cdot}4}{175{\cdot}6}, \quad \text{where } x=(X-\bar{X})$$

$$y=(Y-\bar{Y})$$

$$=0{\cdot}583$$

The regression line is given by—

$$Y-\bar{Y}=b_{YX}(X-\bar{X})$$

or $$Y-6{\cdot}4=0{\cdot}583(X-11{\cdot}4)$$

or $$Y=6{\cdot}4+0{\cdot}583X-6{\cdot}6462$$

or $$Y=0{\cdot}583X-0{\cdot}2462$$

To test the significance of b_{YX}, we take the null hypothesis—

Ho : That its population value is zero *i. e.* $\beta=0$

Now we compute the statistic

$$t=\frac{b_{YX}-\beta}{\text{S.E. of } b_{YX}}, \text{ where S. E. of } b_{YX}=\sqrt{\left(\frac{\text{S. S. Y}-b^2{}_{YX}\,\text{S. S.}_X}{(n-2)\ \text{S. S}_X}\right)}$$

$$=\sqrt{\left(\frac{79{\cdot}6-({\cdot}583)^2\,175{\cdot}6}{13\times175{\cdot}6}\right)}$$

$$\therefore \quad \text{S. E. of } b_{YX}=\sqrt{\left(\frac{79{\cdot}6-0{\cdot}3399\times175{\cdot}6}{2282.8}\right)}$$

$$=\sqrt{\left(\frac{79{\cdot}6-59{\cdot}6864}{2282{\cdot}8}\right)}=\sqrt{\left(\frac{19{\cdot}9136}{2282{\cdot}8}\right)}$$

$$=\sqrt{({\cdot}0009)}={\cdot}03 \text{ approx.}$$

$$\therefore \quad t=\frac{0{\cdot}583-0}{0{\cdot}3}=19{\cdot}43, \quad t_{{\cdot}05}(13)=2{\cdot}160.$$

Conclusion : The calculated value of t is greater than the tabulated value of 't' at 5 percent level for 13 d. f.

Hence the regression coefficient b_{YX} is significant.

Exp. 15. The following table gives the temperature of unhusked rice and the percentage breakage of rice grain in milling—

Temp. (degree) :	33	34	29	35	38	28	29	36	34	30
% breakage :	37	26	26	27	30	24	25	28	30	24

Calculate the coefficient of correlation between the temperature and percentage breakage and test its significance ? State your conclusions ? *(M. Sc. Ag. Agra, 1961)*

Sol. Taking (X) for temperature (in degree) and (Y) for the percentage breakage of rice in milling, we compute the following table for the calculation of the coefficient of correlation—

$A_X = 33$ $A_Y = 27$

S. N.	X	$\xi = X - A_X$	ξ^2	Y	$\eta = Y - A_Y$	η^2	$\xi\eta$
1	33	0	0	37	+10	100	0
2	34	+1	1	26	−1	1	−1
3	29	−4	16	26	−1	1	+4
4	35	+2	4	27	0	0	0
5	38	+5	25	30	+3	9	+15
6	28	−5	25	24	−3	9	+15
7	29	−4	16	25	−2	4	+8
8	36	+3	9	28	+1	1	+3
9	34	+1	1	30	+3	9	+3
10	30	−3	9	24	−3	9	+9
Totals	$n=10$	$-4=\Sigma\xi$	$106 = \Sigma\xi^2$	$n=10$	$+7=\Sigma\eta$	$143 = \Sigma\eta^2$	$56 = \Sigma\xi\eta$

Hence we compute,

$$r=\frac{\Sigma\,\xi\eta-\dfrac{\Sigma\xi\,\Sigma\eta}{n}}{\sqrt{\left\{\left(\Sigma\xi^2-\dfrac{(\Sigma\xi)^2}{n}\right)\right\}}\sqrt{\left\{\left(\Sigma\eta^2-\dfrac{(\Sigma\eta)^2}{\eta}\right)\right\}}}$$

$$=\frac{56-\dfrac{(-4)(7)}{10}}{\sqrt{\left\{\left(106-\dfrac{(-4)^2}{10}\right)\right\}}\sqrt{\left\{\left(143-\dfrac{(7)^2}{10}\right)\right\}}}$$

$$=\frac{56+2{\cdot}8}{\sqrt{(104{\cdot}4)}\sqrt{(138{\cdot}1)}}=\frac{58{\cdot}8}{\sqrt{(14417{\cdot}64)}}=\frac{5{\cdot}88}{120{\cdot}07}=+0{\cdot}49$$

$\therefore \quad r=+0{\cdot}49$

To test the significance of 'r', we take the null hypothesis—

Ho : that the population correlation coefficient $\rho=0$

Then we compute the statistic

$$t=\frac{r}{\sqrt{(1-r)^2}}\times\sqrt{(n-2)}$$

$$=\frac{0{\cdot}49}{\sqrt{1-[(0{\cdot}49)^2]}}\times\sqrt{8}$$

$$=\frac{0{\cdot}49}{\sqrt{(0{\cdot}8799)}}\cdot\sqrt{8}=0{\cdot}49\times\sqrt{\frac{(8)}{0{\cdot}8799}}=0{\cdot}49\times3{\cdot}014$$

$\therefore \quad t=1{\cdot}4769$ and $\underset{\cdot05}{t}(8)=2{\cdot}306$

Conclusion : Thus the calculated value of 't' is less than the tabulated value of 't' at 5% level of significance for 8 d. f. and we conclude that the 'r' is not significant. It means that the temperature (in degree) and % breakage of rice in milling are uncorrelated.

Exp. 16. From observations of no. of bolls (X) and yield of seed cotton (Y) taken on 50 cotton plants, the following quantities were calculated—

$\overline{X}=30, \quad S(X-\overline{X})^2=1225{\cdot}0$

$\overline{Y}=25$ gm. $\quad S(Y-\overline{Y})^2=441{\cdot}0$ gm^2.

$S(X-\overline{X})(Y-\overline{Y})=451{\cdot}0$ gm.

Calculate the regression of yield on boll number and test its significance ? Predict the mean yield per plant of the plants bearing 35 bolls. What would be the regression of boll number on yield ? What is the precise physical meaning of two regression coefficients ?

(M. Sc. Ag. Agra 1958)

Sol. The regression of yield on boll number is given by

$$b_{YX}=\frac{S\,(X-\overline{X})(Y-\overline{Y})}{S\,(X-\overline{X})^2}=\frac{451}{1225}=0{\cdot}3681\approx 0{\cdot}37$$

The regression of boll number on yield is given by,

$$b_{XY}=\frac{S\,(X-\overline{X})(Y-\overline{Y})}{S\,(Y-\overline{Y})^2}=\frac{451}{441}=1{\cdot}0226\approx 1{\cdot}02$$

The regression line of yield on the number of bolls is—

$$Y-\overline{Y}=b_{YX}\,(X-\overline{X})$$

or $\quad Y=25+0{\cdot}37\,(X-30)$

or $\quad Y=0{\cdot}37X+13{\cdot}9$

The mean yield per plant for the plant bearing 35 bolls will be given by—

$$Ye=0{\cdot}37\times 35+13{\cdot}9$$

or $\quad Ye=12{\cdot}95+13{\cdot}9$

or $\quad Ye=26{\cdot}85$ gm.

To test the significance of b_{YX}, we take the null hypothesis—

Ho : that the value of the population regression coefficient is zero (*i. e.* $\beta=0$).

Then we compute the statistic,

$$t=\frac{b_{YX}-\beta}{\text{S. E. of } b_{YX}},\ \text{where S. E. of } b_{YX}=\sqrt{\left(\frac{S.\,S_Y-b^2_{YX}\,S.S_X}{(n-2)\,S.S.}\right)}$$

$$=\sqrt{\left(\frac{441-(0{\cdot}37)^2\times 1225}{48\times 1225}\right)}$$

$$=\sqrt{\left(\frac{273{\cdot}2975}{58800}\right)}$$

$$=\sqrt{(0{\cdot}004647}=0{\cdot}068$$

$$\therefore\quad t=\frac{0{\cdot}37-0}{0{\cdot}068}=5{\cdot}44 \text{ and } \underset{\cdot 05}{t}\,(48)=2{\cdot}01$$

Result : Thus the calculated value of 't' is greater than the tabulated value of 't' at 5% level of significance for 48 d. f. and we conclude that the regression of yield on boll number is significant.

The coefficient of regression of Y on X and of X on Y give estimates of changes of Y and of X respectively for the corresponding unit changes in X and Y.

Exp. 17. What is linear regression ? What principle is employed in obtaining the estimate of 'b' constants in a regression equation ?

For two characters X and Y, it is found that

$$r=+0{\cdot}8 \quad , \quad \sigma_X=4$$
$$\overline{X}=25 \quad , \quad \sigma_Y=5$$
$$\overline{Y}=22$$

Calculate (i) the expected value of X when Y=12
(ii) the expected value of Y when X=33

(M. Sc. Ag. Agra 1962)

(iii) test the significance of b_{YX} for $n=11$.

Sol. When the points on the scatter-diagram seem to cluster about a straight line, it suggests some linear relationship between the variables and this st. line is called *the line of regression.* It gives the average value of the dependent variable corresponding to a given value of the independent variable. When both the variables take the role of independent variable, we have two regression lines—

(i) *the regression line of Y on X :* which gives the average value of Y for a given value of X.

(ii) *the regression line of X on Y :* which gives the average value of X for a given value of Y.

In general, the two lines are different except in the case when the correlation is perfect.

Further, Let $Y=a+bX$, be the equation of the line of regression of Y on X, where a and b are the constants of regression. To estimate these constants a and b, the principle of *least squares* is employed.

The P. L. S. consists in finding the sum of squares of the Y from its estimated value and then minimizing it with respect to a and b. The estimates of a and b thus obtained are—

$$a=\left(\overline{Y}-r\,\frac{\sigma_Y}{\sigma_X}.\,\overline{X}\right)$$

$$b=\frac{\text{cov}\,(X,\,Y)}{V\,(X)}=r.\,\frac{\sigma_Y}{\sigma_X}\,.$$

This estimated value $b=r\frac{\sigma_Y}{\sigma_X}$ is called the regression-coefficient of Y on X and is written more precisely by the symbol b_{YX}. It gives the average change in Y for a unit change in X.

The regression line of Y on X is given by

$$Y-\overline{Y}=b_{YX}(X-\overline{X}),\text{where } b_{YX}=r\,\frac{\sigma_Y}{\sigma_X}=0{\cdot}8\times\frac{5}{4}=1{\cdot}0$$

or $Y-22=1{\cdot}0\,(X-25)$

or $Y=22+X-25$

or $Y=X-3$

The expected value of Y when $X=33$, is given by

(ii) $Ye=33-3=30$ $\qquad\therefore\ Ye=30$

Similarly, the regression line of X on Y is given by

$$X-\overline{X}=b_{XY}(Y-\overline{Y}),\text{where } b_{XY}=r\frac{\sigma_X}{\sigma_Y}=0{\cdot}8\times\frac{4}{5}=0{\cdot}64$$

or $X-25=0{\cdot}64\,(Y-22)$

or $X=25+0{\cdot}64Y-14{\cdot}08$

or $X=0{\cdot}64Y+10{\cdot}92$

The expected value of X when $Y=12$, is given by

(i) $Xe=0{\cdot}64\times12+10{\cdot}92=7{\cdot}68+10{\cdot}92=18{\cdot}60$

$\therefore\ Xe=18{\cdot}60$

$$\text{Hence we have}\quad \left.\begin{array}{l} Y=X-3,\ X=0{\cdot}64Y+10{\cdot}92 \\ Ye=30,\quad Xe=18{\cdot}60 \end{array}\right]\text{..........}\textbf{Ans.}$$

We have computed the regression coefficient $b_{YX}=1{\cdot}0$. To test the significance of b_{YX}, we take the null hyp.

Ho. : That the population value (β) of the regression coefficient is zero.

For this we compute the statistic,

$$t=\frac{b_{YX}-\beta}{\text{S. E. of } b_{YX}},\ \text{where S. E. of } b_{YX}=\frac{\sigma^1{}_Y}{\sqrt{[(n-2)]}.\sigma_X}$$

$$\text{and}\quad \sigma^1{}_Y=\sqrt{(\sigma_Y{}^2-b^2{}_{YX}.\sigma^2{}_X)}$$

$$=\frac{1{\cdot}0-0}{0{\cdot}25}\qquad\qquad =\sqrt{[25-(1)^2\times16]}$$

$$=\sqrt{9}=3.$$

$$\therefore\ t=4{\cdot}0\qquad\qquad \therefore\ \text{S. E. of } b_{YX}=\frac{3}{\sqrt{9\times4}}=\tfrac{1}{4}=0{\cdot}25.$$

But $\qquad t_{\cdot05}(9)=2{\cdot}262.$

Result : The calculated value of 't' is greater than the tabulated value of 't' at 5% level for 9 d. f. Hence we conclude that the regression coefficient b_{YX} is significant.

Exp. (18) : In a partially destroyed laboratory record of the correlation analysis of data, the following results only are legible :

Variance of $X = 9{\cdot}0$,

Regression lines, $8X - 10Y + 66 = 0$

and $40X - 18Y = 214$

Supply (i) the mean values of X and Y,

(ii) the S. D. of Y,

and (iii) the correlation coefficient between X and Y.

(M. Sc. F. 1963)

Sol. (i) We know that the two regression lines intersect at the point $(\bar{X}, \bar{Y})$. Hence, the means can be found out by solving the two regression equations :

$$\left.\begin{aligned} 8\bar{X} - 10\bar{Y} &= -66 \\ 40\bar{X} - 18\bar{Y} &= 214 \end{aligned}\right]$$

or
$$\left.\begin{aligned} 40\bar{X} - 50\bar{Y} &= -330 \\ 40\bar{X} - 18\bar{Y} &= 214 \end{aligned}\right] \text{subtracting we get}$$

$$-32\bar{Y} = -544 \quad \text{or} \quad \left.\begin{aligned} \bar{Y} &= 17 \\ \text{and} \quad \bar{X} &= 13 \end{aligned}\right]$$

(iii) Among the two equations, we are not certain which one of them is the regression of Y on X and the regression of X on Y. There may be two possibilities :

(a) $8X - 10Y + 66 = 0$ may be the regression line of X on Y.

or $8X = 10Y - 66$,, ,, ,, ,, ,,

and $40X - 18Y = 214$ may be the regression line of Y on X.

or $18Y = 40X - 214$,, ,, ,, ,, ,,

Then $b_{XY} = \frac{10}{8}$ and $b_{YX} = \frac{40}{18}$

So $b_{XY} . b_{YX} = r^2 = \frac{10}{8} \times \frac{40}{18} > 1$, which is not possible.

(b) $8X - 10Y + 66 = 0$ may be the regression line of Y on X

or $10Y = 8X + 66$,, ,, ,, ,, ,,

and $40X - 18Y = 214$ may be the regression line of X on Y

or $40X = 18Y + 214$. ,, ,, ,, ,, ,,

From these regression lines, we have

$$b_{YX} = \frac{8}{10} \text{ and } b_{XY} = \frac{18}{40}.$$

Hence $b_{YX} . b_{XY} = \frac{8}{10} \times \frac{18}{40} = 0{\cdot}36 < 1$, hence this set of regression equations is possible.

Now $\quad r^2 = 0{\cdot}36$

or $\quad r = \pm 0{\cdot}6$,

But we take $r = +0{\cdot}6$ as the sign of the two regression coefficients are the same. Hence the correlation between X and Y is +ve.

(ii) We know that $\quad b_{YX} = r \dfrac{\sigma_Y}{\sigma_X}$

or $\quad \dfrac{8}{10} = 0{\cdot}6 \dfrac{\sigma_Y}{3}$

$$\therefore \quad \frac{8 \times 3}{10 \times 0{\cdot}6} = \sigma_Y$$

$$4 = \sigma_Y .$$

Hence we get,

$$\left.\begin{array}{l} \bar{X} = 13, \ \bar{Y} = 17, \\ \sigma_Y = 4, \ \ r = +0{\cdot}6 \end{array}\right] \text{......}\textbf{Ans.}$$

EXERCISE 5.

Q. (1) Define the correlation coefficient ? Find the correlation coefficient for the following values of X and Y—

X : 8, 10, 15, 17, 20, 22, 24, 25
Y : 25, 30, 32, 35, 37, 40, 42, 45

Q. (2) What are the regression lines ? Give their point of intersection.

From the following data, calculate (a) the regression coefficient of monthly yields of guinea grass on age of grass, (b) the regression line of monthly yields of guinea grass on age of grass ?

Age (X) : 1, 2, 3, 4, 5, 6, 7, 8
Monthly yield (Y) : 85, 73, 65, 41, 20, 36, 16, 24

Q. 3. Give the relation between the correlation coefficient and the regression coefficients ?

The following table gives the heights and dry weights of 17 plants, calculate

(i) the coefficient of correlation,
(ii) the regression coefficient of weights on heights,
and (iii) the regression line of weights on heights.

Height (X inches): 50 44 54 46 64 37 77 62 45 69 44 57 63 68 45 71 50
dry wts. (Y gms.): 44 25 44 26 41 21 60 52 15 55 42 58 37 36 40 58 36

Q. (4) What do you understand by perfect positive and perfect negative correlation ? Is a significance test needed for this value ? The correlation coefficient between *days to head* and *days to mature* for 25 varieties was found to be 0·4, is this significant ? *[M. Sc. Ag. Agra, 1962]*

Q. (5) Explain the term regression equation ? Why should there be two regression equations ? Write down the two regression equations that may be associated with the following pairs of values—

X : 152, 114, 138, 154, 144, 153, 141, 117, 156
Y : 193, 300, 414, 594, 676, 549, 320, 483, 481 (*I. A. S. 1951*)

Q. 6. The ranks of the same 16 students in two subjects A and B were as follows—(Two numbers within brackets denote the ranks of the students in A and B respectively)

(1, 1) (2,10), (3,3), (4,4), (5,5), (6,7), (7,2), (8,6) (9,8), (10, 11), (11, 15) (12,9) (13,14) (14,12) (15,16) (16, 13)

Calculate the rank correlation for proficiencies of this group in subject A and B ? *[M. Sc. Agra, 1952]*

Q. 7. A study of wheat prices at Hapur and Karanchi yields the following data—

	Hapur	Karanchi
Average price/maund :	2·463	2·797
Standard deviation :	0·326	0·207

The coefficient of correlation between two prices is +0·774.

Estimate from the above data, the most likely price of wheat—

(i) at Hapur, corresponding to the price of Rs. 2·334 at Karanchi.

(ii) at Karanchi, corresponding to the price of Rs. 3·052 at Hapur.

Q. 8. For two characters X and Y,

$$\bar{X}=30,\quad \bar{Y}=7{\cdot}143,\quad \Sigma(X-\bar{x})^2=1008,$$

$$\Sigma(Y-\bar{Y})^2=11{\cdot}017,\quad \Sigma(X-\bar{X})(Y-\bar{Y})=102{\cdot}0,$$

calculate the coefficient of regression of Y on X and test its significance ?

Q. 9. (a) Find the coefficient of correlation between **X** and Y, and their means from the following regression lines :

$$10Y-16X-21=0$$

and $$2Y-5X+16=0$$

(b) Plot the lines on the graph and give your view about the correlation ?

(c) From the graph obtain $\bar{X}$, $\bar{Y}$ and compare them with those of (a).

Q. 10. Find the most likely price in Bombay corresponding to the price of Rs. 70 at Calcutta from the following data :

Average price at Calcutta	65 Rs.
,, ,, ,, Bombay	67 Rs.
S. D. at Calcutta	2·5 Rs.
,, ,, Bombay	3·5 Rs.

Coefficient of correlation is +0·8 between the two prices of the commodity in the two cities. *(M. A. Agra, 1954)*

Q. 11. From the following data, find out the two regression coefficients ?

$\Sigma\xi=-1,\ \Sigma\xi^2=229,\ \Sigma\eta=-4,\ \Sigma\eta^2=92,\ \Sigma\xi\eta=139,$
$n=6$, where ξ, η are the deviations from the assumed means.

Also test the significance of the coefficients ?

Q. 12. (a) In the calculation of regression coefficient for 18 pairs of observations, it was found that $b_{YX}=1{\cdot}12$ and S. E. $b_{YX}=0{\cdot}06$. Is the estimate significantly different from the expected value $1{\cdot}0$.

(Hint : take $\beta=1{\cdot}0$)

(b) For 8 pairs of observations of X and Y, the following were calculated :

$$\Sigma X=360,\quad \Sigma Y=36,\quad \Sigma XY=1210,\quad \Sigma X^2=20988,$$
$$\Sigma Y^2=204.$$

Find out b_{XY} and test its significance ?

Q. 13. The following table enumerates the marks obtained by a class of students in statistics in 1st. and 2nd. paper :

Marks in 1st paper (X) :

80, 45, 55, 56, 58, 60, 65, 68, 70, 75, 85.

Marks in 2nd paper (Y) :

82, 56, 50, 48, 60, 62, 64, 65, 70, 74, 90.

Calculate (i) The regression coefficients ?

(B. Sc. Agra, 1956)

(ii) Obtain the scatter diagram and comment on correlation between X and Y.

Q. 14. A sample of paired variates is given below :

X_1 : 9 8 7 7 6 5 3 3 1 1

X_2 : 9 9 8 6 6 5 4 3 1 1

(a) Calculate the correlation between the two variables and interpret the same ?

(b) Illustrate the scatter of the points in a diagram ?

(M. Sc. Ag. Agra, 1965)

(c) In a sample of 20 pairs of values, total solids and fat contents gave the correlation coefficient of $0{\cdot}6$. What would you conclude from this ?

(M. Sc. Ag. Agra, 1959)

ANSWERS

(1) $r=+0{\cdot}98$

(2) (a) $b_{YX}=-9{\cdot}762$, (b) $Y-45=-9{\cdot}762\,(X-4{\cdot}5)$

(3) (i) $r=0{\cdot}7306$ (ii) $b_{YX}=0{\cdot}8394$
(iii) $Y-40{\cdot}59=0{\cdot}8394\,(X-55{\cdot}65)$

(4) $t=5{\cdot}23$

(5) $Y-444{\cdot}45=2{\cdot}408\,(X-141)$
and $X-141=0{\cdot}025\,(Y-444{\cdot}45)$

(6) $r=0{\cdot}8$

(7) (i) 1·899 Rs. (ii) 3·086 Rs.

(8) $b_{YX}=0{\cdot}101$, $t=732$ approx.

(9) (a) $r=+0{\cdot}8$, $\overline{X}=11{\cdot}22$ approx., $\overline{Y}=20{\cdot}05$

(10) $Y=72{\cdot}6$ Rs.

(11) $b_{YX}=0{\cdot}60$, $b_{XY}=1{\cdot}55$, $t=7{\cdot}5$, $t=7{\cdot}75$.

(12) (a) $t=2$, (b) $b_{XY}=-9{\cdot}762$, $t=5{\cdot}5$.

(13) (i) $b_{YX}=0{\cdot}99$, (ii) $b_{XY}=0{\cdot}85$.

(14) (a) $r=+0{\cdot}98$, (c). $t=3{\cdot}18$.

Appendix I & II

Tables & A. U. Papers

‘t’—Table (values at 5% and 1% levels)

Probability / D.F. (ν)	.05	·01
1	12·706	63·657
2	4·303	9·925
3	3·182	5·841
4	2·776	4·604
5	2·571	4·032
6	2·447	3·707
7	2·365	3·499
8	2·306	3·355
9	2·262	3·250
10	2·228	3·169
11	2·201	3·106
12	2·179	3·055
13	2·160	3·012
14	2·145	2·977
15	2·131	2·947
16	2·120	2·921
17	2·110	2·898
18	2·101	2·878
19	2·093	2·861
20	2·086	2·845
21	2·080	2·831
22	2·074	2·819
23	2·069	2·807
24	2·064	2·797
25	2·060	2·787
26	2·056	2·779
27	2·052	2·771
28	2·048	2·763
29	2·045	2·756
30	2·042	2·750
∞	1·960	2·576

EDUCATIONAL PSYCHOLOGY QUESTIONS
Occurred in A. U. B. T. EXAMINATIONS

1. Discuss the meaning and importance of measures of central tendencv and dispersion, with special reference to their use in the handling of a class. (A. U. B. T., 1960)

2. Calculate coefficient of correlation from the data given below by Spearman's rank method and interpret the result :—

		Marks in Eng.	Marks in Maths.
A	...	22	32
B	...	27	39
C	...	28	40
D	...	15	20
E	...	32	21
F	...	14	21
G	...	7	16
H	...	9	35
I	...	5	48
J	...	9	16

(A. U. B. T., 1961)

3. Explain fully the meaning and importance of measures of central tendency and dispersion, with special reference to their use in the handling of a class. (A. U. B. T., 1962)

4. The following are the scores in English and Mathematics of ten students. Calculate the coefficient of correlation and comment upon the result.

	A	B	C	D	E	F	G	H	I	J
English ...	25	27	30	40	37	15	14	60	80	10
Mathematics ...	24	29	35	60	45	10	13	50	70	21

(A. U. B. T., 1962)

5. What is standard deviation? Calculate the standard deviation in the following distribution :—

Units	Frequency
0–9	10
10–19	20
20–29	20
30–39	50
40–49	40
50–59	15
60–69	4

(A. U. B. T., 1963)

6. What do you understand by 'Coefficient of Correlation'? Calculate the degree of closeness between English and Mathematics in the following distribution of marks of eight students in a class—

	English	Mathematics
A	60	10
B	50	6
C	50	8
D	40	8
E	60	7
F	30	4
G	35	5
H	50	7

(A. U. B. T., 1964)

7. The I Q and scores on handwriting test of 25 pupils are given below. Calculate the coefficient of correlation between them and interpret the result. Use rank-difference method.

Scores on handwriting test :— 75, 58, 55, 50, 40, 62, 57, 53, 49, 38, 61, 57, 53, 48, 36, 59, 56, 51, 47, 35, 59, 56, 50, 46 and [illegible]4.

I Q :—118, 108, 107, 102, 100, 122, 115, 115, 100, 97, 119, 112, 112, 103, 93, 110, 109, 109, 95, 89, 110, 108, 108, 98 and 90.

(A. U. B. T., 1959)

8. Discuss the more important method of assessing personality, with special reference to their validity and reliability.

(A. U. B. T., 1960)

9. What are group tests of intelligence? How is the intelligence of a group of children assessed through them? What use can the school make of their results for educational purposes?

(A. U. B. T., 1961)

Form No. 3.

PSY, RES.L-1

Bureau of Educational & Psychological Research Library.

The book is to be returned within the date stamped last.

---

26 JUN 1973

WBGP-59/60-5119C-5M

1969